THE DEVELOPMENT AND EVALUATION OF A SYSTEMATIC TRAINING PROGRAM FOR INCREASING BOTH RATER RELIABILITY AND RATING ACCURACY

BY

SO YOUNG JANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Fred Davidson, Chair
Professor Hua-hua Chang
Associate Professor Jinming Zhang
Assistant Professor Randall Sadler

Abstract

The primary purposes of this study are to identify the characteristics of modeling a rater training program and to develop an efficient training model at the University of Illinois at Urbana-Champaign. This study focuses substantially on a basic conception of rater reliability including true score measurements of examinees' language proficiency. This study was conducted based on a definition of rater reliability achieved by the reinterpretation of the various meanings of reliability.

For these purposes, a basic framework of standardization was achieved using training theories, and this study proposes that a rater training program can be standardized by accomplishing innovative systematic changes that consider (a) the relevant literature, (b) the test instrument itself, (c) the test procedure, and (d) contextual effects such as the characteristics of the stakeholders, their concerns, the structure of the test, or washback effects of the test use.

This study utilized a modified version of Lynch's program evaluation model (1996; 2003) to collect evidence from different sources, including data drawn from the entire evaluation process ranging from needs analysis to a feedback system based on the final product of the evaluation. The effectiveness of both the training program and the individual performances were identified by incorporating all sources of data collected using measurement theory.

Mixed methods were proposed for the data analysis. The data analysis involved an investigation of training effectiveness by measuring raters' scoring reliability, and providing a new training program for raters' professional improvement. Quantitative data analysis was proposed for analyzing the surveys, the rating corpus, and training effectiveness. Qualitative and document analysis were also essential for analyzing relevant training materials and workshop observation as well as exploring the degree of change in the perceptions of the raters.

The results of this study provide educational implications for language testing. At the program level, standardized training contributed to shared responsibilities among test users. The results support the idea that the professionalism of the raters could be improved by providing access to similar quality input which can reinforce their learning and skills via training.

The salient value of this dissertation is the collaboration with stakeholders in a test administration situation. Stakeholders' concerns and challenges were clearly identified, shared, and resolved with the practitioners (the EPT trainer and raters). In addition, I recognize the importance of a balance between understanding fundamental theoretical underpinnings and applying theory through practical experience. It could be concluded that this study contributes to the enhancement of rating validity and the cumulative growth in scoring reliability, as well as a positive washback effect for the future rater training program.

To my father and mother, my brother, Won-seok, and my sister, Jung-hwa

Acknowledgments

I would like to give thanks to my great committee members, colleagues, friends and my family in Korea. First of all, this study could not have been completed without the committee members' guidance and their support. In particular, I greatly appreciate Dr. Fred Davidson, a dissertation director as well as my advisor. I could not forget his professional support and encouragement whenever I confronted challenges and difficulties during my studies. He is my academic role model and my mentor. Through his class, I found my identity and my philosophical foundation as a language tester, a researcher and a language teacher and it helped me to move forward in my research and my graduate life in Champaign.

I would like to thank to my committee members, Dr. Chang, Dr. Sadler, and Dr. Zhang. Their guidance and valuable comments allowed me to re-consider substantial issues in my dissertation topic and were helpful to enhance the quality of my dissertation.

I would like to express my sincere gratitude to my research participants for their help. In my dissertation, many people participated in the different research phases. Particularly, I appreciate Prof. Susan Ahmed, a coordinator of the ESL writing courses. Her professional knowledge and experience was helpful in refining my study. I had difficulty in recruiting raters for the post-rating session at the final stage of my research and I could only complete my dissertation with the help of ESL teacher raters Heejeong, Hyunbin, Jing-Ru, Heejung, Jongyeon and Qishu who participated in the post-rating session. I also appreciate the help of Sun-joo, the EPT G.A, and I was happy to collaborate with her during the research.

I will not forget my experience as a research assistant at the Foreign Language Assessment Group (FLAG). This research experience has provided me an opportunity to apply my theoretical knowledge to practical concerns and motivated me to continue my study. I would

like to thank my colleagues Kaddessa, Ji-young, Youngshin, Huei-Lien (Tammy), Chih-Kai (Cary), and Kyeongsuk.

I should also mention my lovely friends, Eun-young, Ji-young, Cherly, Eun-kyung, Ki-ryung, E-jung, Malsook, Chaeyoon, Ha-young, Eun-ae, Youngju, Sookyung, Eunhyun, Eugene and Sekyung. Moreover, I greatly appreciate Dr. Shin, my former advisor in my master's program in Korea.

Finally, I would like to dedicate this dissertation to my family who allowed me to finish this long journey. Words cannot express my appreciation to my mother and my father who are my spiritual mentors. During last seven years, my mother has prayed to God every day for my successful completion of doctoral life in Champaign. I also thank my father, my brother, Wonsuk, and my sister Junghwa. I did not do anything here without my family's emotional and financial support. I have decided to return to my home country after this, and I am really excited to have new life as a researcher and as a family member in Korea.

Table of Contents

List of Tables

List of Figures

Chapter One

Introduction

Statement of the Research Problem

During the past 30 years, much research on rater reliability in the area of language testing has been conducted (Bachman, Lynch, & Mason, 1995; Brown, 1995; Charney, 1984; Cumming, Kantor, & Powers, 2001; Erdosy, 2004; Freedman & Calfee, 1983; Lumley, & McNamara, 1995; McNamara, 1996; Shi, 2001; Shin, 2001; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994a; 1994b; 1998). Most of these studies have focused on scoring variability as it relates to rater background. Some studies have explored the cognitive scoring process of raters, and some have investigated issues of rater training.

However, this study assumes a program-evaluative view of rater screening and training. This is important because a more systematic training program makes raters more reliable via analysis of their rating patterns and practices as well as an understanding of the nature of the rating process. The investigator will cooperate with the administrators of a test (The ESL Placement Test at University of Illinois at Urbana-Champaign, hereafter EPT) to set up the goals of the program evaluation. Sources of variability and ways to enhance scoring reliability and validity will be considered from a wider perspective, particularly in terms of the effectiveness of the rater training program.

Four perspectives present why professional rater training workshop is necessary for the EPT writing raters (Fulcher & Davidson, 2007; Mathison, 1992; Phillips, 1997). First, operationalized training workshop should be evaluated. Although the EPT provides

operationalized training programs for the writing tests, there is currently no opportunity to evaluate how well the training workshops work and/or what components could be improved. Second, professional training workshop should be standardized. Currently, the EPT training workshop was designed to deliver simple knowledge about rating procedures, rather than enhancing score reliability. Various materials for the workshop, different training methods for the effective delivery of content, and various measures for the improvement of score reliability should be considered through the standardization of training workshop. Third, professional training workshop should consider stakeholders' needs, particularly raters' concerns which are currently occurring in rating context. It can encourage raters to become active learners and practitioners who appropriately establish their understanding on scoring process. In addition, it can be easily integrated their knowledge they received from the workshop into rating practice. Finally, professional rater training workshop should be iterative on-going process. A one or two hours training workshop is conducted once at the beginning of the semester; as a result, raters get training once a year, which seems insufficient to address their concerns about scoring processes. Continuous support is necessary for improving raters' professional development. Systematic rater training workshop is a method to satisfy these four requirements listed above.

Significance of This Study

This study will focus on development of Systematic training program for the EPT essay raters. Most of the recent research about rater reliability has focused either on a discrete single aspect by providing a single statistical index, or on one aspect of raters'

individual backgrounds, such as raters' background information, interaction between rater and task, or raters' decision making processes, based on a summary of interview or verbal protocol/think aloud methods. It is not easy to find meaningful implications from these studies for the actual improvement of rater reliability. It is suggested by many studies that developing a systematic training program through iterative evaluation could be a way to reduce the variability in rating among raters (Choi, 2002; Shin, 2001; Shohamy et all, 1992; Weigle, 1998). In spite of this suggestion, theoretical frameworks and practical guidance for rater training have had little open discussion, having been handled primarily as an internal practice of individual testing agencies, despite the demand for systematic training programs for enhancing rater reliability. It seems that standardization would be an effort to see scoring problems as a matter of educational system rather than individual responsibility. This study is designed to overcome the lack of theoretical framework to meet practical demands by discussing the possibility of the standardization of rater training programs.

Purpose of This Study

The primary objectives of this study are to develop a systematic screening process for the rater training program of the ESL placement test (here after EPT, essay writing tests), with a fuller perspective of rater reliability. This study will be conducted based on a definition of rater reliability achieved by the reinterpretation of the various meanings of reliability, many of which are having been ignored; the goal here is to return to a basic conception of reliability, namely, that we are substantially interested in true score measurements of examinees' language proficiency. For this purpose, this study proposes

a rater training model for the process of standardization by reviewing the relevant literature; considerations about the test instrument itself; the use of the test procedure; participants' educational training; and contextual effects, such as the characteristics of the stakeholders, the structure of the test battery, or washback effects of the test instrument use, have been fully taken into account. This topic will take a fundamental approach to understanding the entire rating process regarding rater reliability by accomplishing innovative systematic changes via the standardization of the test instrument.

Rationale of This Study

Given these concerns about the EPT, a rater training program should be carefully designed. In line with this, standardization of rater training programs, as suggested by Fulcher and Davidson (2007), seeks to improve scores or rater reliability by considering the appropriate use of the test instrument, standardization of administration procedures. With respect to score reliability, consistent measures within and between raters and the standardization of a rater training program have become central issues in language testing.

Standardization can be achieved and evaluated by following modified Lynch's program evaluation model (1996; 2003) to formulate a basic framework of standardization, which includes the entire evaluation process from needs analysis to feedback system on the basis of the final product of the evaluation. Measurement theory will contribute to evaluating the program's effectiveness and individual raters' performances, including the consistent identification of true score for prototyping with respect to the concept of true score theory.

Concerns for both the effectiveness of a training program and the individual performance will be identified by comparing post-rating results of workshop and control groups. The content and guidance of the training workshop will be provided, based on needs analysis, and interim outputs by collaborating with an EPT administrator. In addition, in terms of the ethics of test use, standardization would help ascertain that the test instruments are used appropriately and that the score as an outcome of the test would be interpreted and justified both statistically and semantically, with some regard to the match between these interpretations. This study suggests that a more systematic theoretical model for designing a training program for assessors should be established, since a better system would assure better outcomes in rating, and a well-organized screening system would enhance rater-reliability.

Background of This Study

The ESL Placement Test (EPT) at the University of Illinois at Urbana-Champaign (UIUC) is a test for placing new international students into the appropriate levels of ESL classes (http://www.linguistics.illinois.edu/students/placement/). It measures two different components of language ability: speaking and writing. This study, however, focuses on the rating process of the writing portion, because the oral test is not a full oral interview, and it is a highly targeted screening measure for a particular ESL pronunciation class. It may, therefore, be misleading to explore standardization of that particular workshop program, because the findings would not generalize to other settings. The written test, on the other hand, shares many characteristics with other writing tests,

which is why the focus of this study is on the standardization of the rater training program for the EPT essay test.

Chapter Two

Literature Review

In this chapter, three major agendas will be discussed regarding new directions of systematic training programs. First, the nature and necessity of standardization will be explored via discussion of the benefits of standardization, the effectiveness of rater training, and the reliability of raters. Secondly, the theoretical and practical features which should be considered in order to move toward the standardization of a rater training program will be discussed. Finally, the realization of standardization in a practical situation, including which parts of the training program should be standardized, will be considered

The Nature of the Standardization of a Rater Training Program

Issues of current training programs. A training program and rating guidelines should be carefully designed to reflect an understanding of the complexity of a particular audience and assessment context, in order to bring out a change in raters' perceptions, workshop practices, trainer and trainee experiences, and the program as a whole. The training workshop plays a critical role in connecting the theoretical constructs of language performance with the operational constructs of the practical situation. In other words, training helps raters use the test instrument appropriately. Through the training workshop, raters should understand the test constructs, the test purposes, and the test procedures.

Secondly, the training workshop should offer opportunities to monitor individual raters' performances by having group activities and feedback sessions. An important

function of a training program is to screen whether a rater is qualified. However, it is true that more empirical research is necessary for answering the fundamental problems of current training programs. Several studies (Bachman, 1988; Bachman & Savignon, 1986; Salaberry, 2000) have criticized the validity and reliability of the ACTFL OPI and the ACTFL OPI Tester Training Manual (hereafter ACTFL-TTM). The ACTFL-TTM was revised in 1999, but they still cannot avoid the criticism that more effort should be made to develop a theoretical model for testing and training content based on empirical data. Salaberry (2000) and Shohamy (1990) stated that the ACTFL-TTM does not reflect the essential features of theoretical and operational models to train interviewers/raters by simplifying information which the guidelines should contain.

Therefore, the training program should be standardized in order to acquire a more systematic rater training program. To accomplish this, more objective standards and methods for evaluation, and regular training, including the capability of performance review and giving and receiving feedback should be required. Training is needed to increase raters' awareness of their problems, and also of the solutions to those difficulties. The terms "standardized" and "systematic" are conceptually almost interchangeable in this study, but the two terms can be distinguishable. Standardization can be defined as the series of activities or procedures which make a rater training program more systematic. A "systematic" training program can be defined as the final product of standardization. In this study, these two terms will be used interchangeably. The training workshop should be a place where raters and trainers can discuss concerns and context-sensitive issues, as well as general principles of institutional policy with regard to the characteristics of the

test takers and the matter of score interpretation, which come up during the scoring process.

Training effectiveness and standardization of a training program. Because of uncertainty about the format of a test instrument and the stakeholders involved in the rating environment, training programs, both onsite and online, are not always successful at enhancing test validity and reliability (Elder, Barkhuizen, Knoch & Randow, 2008). It seems that an effective training program could be achieved through standardization, because standardization can provide an equal quality of training materials and practices. This provision can in turn reduce individual differences. To enhance the effectiveness of a training program, a more systematic and formulated training program is necessary. The goal of standardization is to institute an entire rating system, including rater training, and to promote rater professionalism with institutional support, rather than assigning the responsibility on low score reliability to individual raters (Fulcher & Davidson, 2007).

Standardization for maximizing effectiveness can be accomplished in two areas: a change in a training system through the standardization of rating procedures (instruments) and training inputs (materials and practices); and a change in individual perception by providing access to the standardized materials and program (Weiss, 1998). Moreover, the professionalism of the test users would be enhanced by providing more systematic training on the use of the test, which might positively affect the score interpretation process and improve validity in the end.

Figure 1 exhibits the overall theoretical framework for this study with the three supporting theories for realizing the standardization of a rater training program. In order to reflect multiple perspectives, three major theories (evaluation theory, training theory,

and language testing theory) will serve for designing the systematic rater training program. First of all, program evaluation theory provides a systematic framework for the overall research procedure as well as a philosophical background. It can be adopted in order to see the overall logic of the evaluation process of the training program. Particularly, Lynch's program evaluation theory in applied linguistics, which refers to the context adaptive model, provides guidelines about what should be evaluated and how we should approach the evaluation of each phase (Lynch, 1996; 2003).

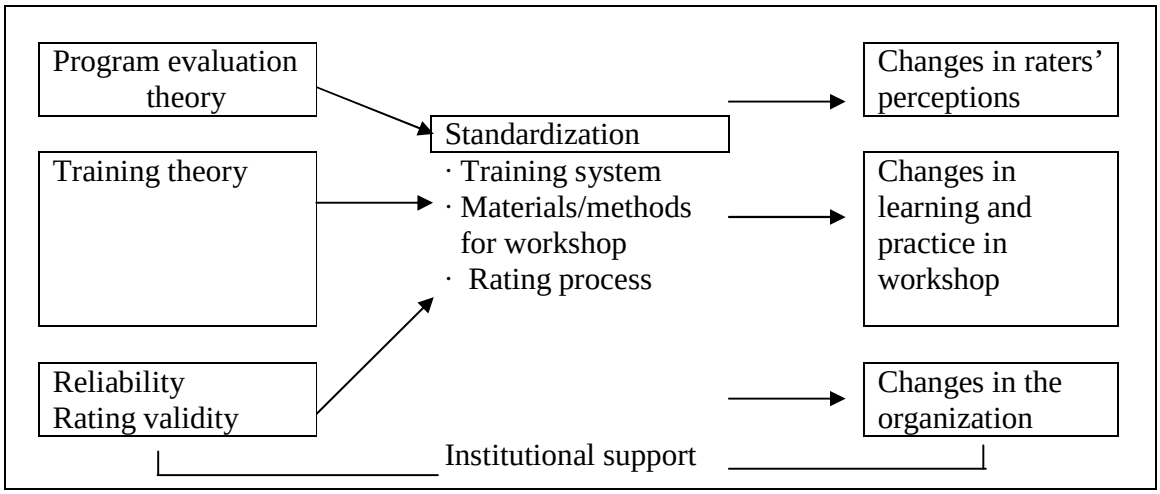


Figure1. Theoretical framework for standardization.

Secondly, training theory, originally adopted from the human resources field, provides a link between the theoretical issues and practical concerns. For the development of a systematic training model to fit the rating context, theoretical training models, specific goals, and methods of the training program will be defined for a particular training program. On the basis of training theory analysis, the most appropriate training materials and methods will be created in cooperation with the EPT trainer and staff.

Next, some issues of rater reliability, measurement, and rating validity will be discussed in combination with the effects of the rater training program. In the area of language testing, much research has been conducted analyzing rating procedures, issues of rating validity, rater reliability, and rater training itself. These are still some of the main topics of research in language testing.

Finally, standardization can be formulated with regard to both theoretical arguments and practical concerns based on empirical data. Bernardin and Buckey (1981) discussed the fact that training effectiveness should be evaluated based on multiple sources in terms of training design (modeling), training methods, measures of rating error, rating scale descriptors, and the nature of rating tasks including the characteristics of the examinees.

In addition, these features, which might affect the procedures of the rating, should be standardized through an appropriate consensus process. The systematic training program should make the rating focus clear by presenting and sharing accurate information (Borman, 1977; 1978; 1979). This strong theoretical background and the practical concerns from the stakeholders' perspectives will complement each other in making changes to raters' perceptions, workshop practice, and the whole rater screening procedure at the institutional level. In order to maximize the effectiveness of the training program, it might be necessary to take into account these substantial features which might affect the quality of the assessors' performance.

Reliability for rater proficiency as well as rating validity. In the tradition of classical testing theory, one reliability index from the psychometric perspective is rater reliability (Anastasi & Urbina, 1997; Bachman, 2004). Reliability index, in other words,

indicates a measurement of error quantified in rater performance. Moss (1994) referred to the conceptual definition of reliability in terms of measurement:

> Theoretically, reliability is defined as "the degree to which test scores are free from errors of measurement….Measurement errors reduce the reliability (and therefore the generalizability) of the score obtained for a person from a single measurement"(AERA et al., 1985,p19). Typically, reliability is operationalized by examining consistency, quantitatively defined, among independent observations or sets of observations that are intended as interchangeable-consistency among independent evaluations or readings of a performance, consistency among performances in response to independent tasks, and so on. (p.6)

Generally, rater reliability can be divided into two types in language testing: inter-rater reliability and intra-rater reliability. Intra-rater reliability measures how consistently raters maintain their own rating severity, as Moss stated above. It is highly possible that raters with low intra- rater reliability maintain inconsistent rating because of errors or other variables. Lack of rating consistency is related to low predictability of rating patterns as to how severely or leniently the rater applies his/her rating standards to examinees (Bachman, 1990).

In comparison to intra-rater reliability, inter-rater reliability is the degree of agreement among raters. It does not really imply overall reliability but rather simply shows common patterns of a rater's severity or leniency, characteristics in judgment making. Inter-rater reliability is less meaningful, because severity can be easily adjusted if consistency is insured (Bachman & Palmer, 1996; McNamara, 1996).

In this study, however, score reliability during training can be refined by elaborating on the traditional concept of reliability. Although reliability is displayed with one single statistical index, it implies a variety of meanings of reliability. Mislevy (as cited in Moss, 2004) described the different properties of reliability in his article:

> characterized four different senses of " reliability": (a) true score reliability, as reflected in classical test theory; (b) reproducibility, as reflected, for instance, in "proportions of agreement among raters, decision-consistency coefficients, and generalizability coefficients (Mislevy, 1994, p.6), (c)differential likelihood, as reflected, for instance, in item response theory, and more generally, in probability based reasoning, "where the relative likelihood of an observation under alternative 'true states' is the weight of evidence it provides for each" (Mislevy, 1994, p.6), and finally (d) credibility, as used in "common parlance," where "reliability" simply means the extent to which information can be trusted . (Mislevy,1994, p.8)

Mislevy described the several different properties of reliability in addition to the basic concept of "consistency". In measurement, true score theory can be explained that the true score can be defined as the expected observable score. The observable score (an examinee's ability) consists of two components: true ability, and random error. In other words, high score reliability can be achieved by exploring what exactly the true score is, and what measurement error is although the nature of a true score conceptually is unknown (Allen & Yen, 1979; Goodenough, 1950; Lord & Norvick, 1968).

Secondly, reliability has the property of reproducibility. For instance, scores measuring the same performance (same true score) within two different time periods show consistency if the scores show reliability across the measurement contexts. This refers to intra-rater reliability, but also, this discussion could be extended to agreement among raters (inter-rater reliability). Scores should be able to be reproduced between raters. This shows that agreement is also an essential feature in reaching high score reliability. Henning (1987) states:

> With a group of examinees, reliable measurement is indicated by a tendency to rank order the entire group in the same way on repeated administrations of the test. (p. 73)

Moss (1994) stated that both intra- and inter-rater reliability are indices of reliability or generalizability. Inter-rater reliability can be seen as an important factor in

determining rating accuracy by ensuring agreement among raters. Schaefer (2008) suggested that low agreement among raters could produce a misfit rating pattern. Weigle (1994a), however, discussed inter-rater reliability with from a different viewpoint, asserting that it is no indication of rater reliability, because agreement between raters tends to increase or decrease depending on the rating partner or trainer when raters used the same assessment criterion.

DeRemer (1998) was concerned about a more substantial meaning of a score — specifically, what high statistical agreement really means in the interpretation of a score. Although raters assign the same score to the examinee, it is possible that the agreement might be superficial. However, we can obtain from inter-rater reliability information about how useful the rater training is (Borman, 1978), because agreement could be a good index for rating accuracy or validity. In order to expand the extent of variance, error variance includes not only within group variance, but between group variance. A high inter-rater reliability is a benchmark for a successful rater-training program.

Finally, credibility is part of the concept of reliability. According to Mislevy (2004), credibility refers to the degree to which "information can be trusted and it seems to reflect a semantic interpretation of true scores beyond the mathematical interpretation. Credibility seems to refer to the accuracy of information, such as the accuracy or validity of rating results. Parkes (2007) considers the complexity of reliability:

> So what exactly is a reliability argument? A reliability argument would have these six components: 1. A determination of the social and scientific values of dependability, consistency, accuracy, etc. most relevant to the scenario at hand. 2. Clear statements of the purpose and the context of the assessment. 3. The definition of a replication in the particular assessment. 4. A determination of the tolerance or level of reliability needed. 5. The evidence. 6. The judgment: Pulling it all together. (p.6)

For the purposes of rater training, some different approaches toward reliability will be shown, taking into account all possible considerations of a rating environment. A single reliability index alone seems to be insufficient for judging an expert rater, and for finding implications for the enhancement of the quality of rating via the training session. Bernardin and Walter (1977) showed that rating quality might be influenced not only by the training format but also by the quality of rating scales. Borman's (1979) and Weigle's (1994a; 1994b) findings also support Bernardin's and Walter's results. Although his training was successful for reducing errors, rating accuracy was not satisfied because of personal preference. Lumley (2002) concerned validity of rating. Therefore, the definition of rater reliability needs to be expanded to mirror the entire scoring process by taking into account the rating environment at both the individual level and the educational program level.

In this study, four aspects of reliability—consistency, rating accuracy, rater agreement and rating validity—were considered. The definition of consistency and agreement followed the traditional concepts of measurement theory. Rating accuracy was considered by comparing with official scores (or trainer's scores). Accuracy is slightly different from agreement (inter rater reliability); agreement can be estimated in terms of peer rater severity. Finally, rating validity can be evaluated with a more qualititative approach to looking at raters' reflections or group discussions, and it can be also determined by incorporating quantitative measures of rating performance.

Effectiveness of standardization on the reduction of measurements errors. Bachman & Savignon (1986), and Bachman (1988) discussed the effect of the test method or instrument on errors, and finally on the reduction of reliability. It was shown

that standardization of the test procedures would be one possible way to reduce systematic and random errors coming from the test method (Bachman & Savignon; 1986).

> One way to minimize random error is to standardize the test method, or to control the conditions under which the test is given so that they are the same for all tests. In an oral interview, standardization typically involves carefully training interviewers so that they:1) follow the same elicitation procedures; 2) use the same definitions for rating the performances of different individuals; and 3) administer the interview in the same environment and in appropriately the same amount of time. However, the problem created by standardization is that controlling these method factors causes them to have a systematic effect on test scores. (p384)

However, Bachman and Savignon (1986) discussed a dilemma regarding measurement errors and standardization of the test method. For example, systematic errors may be caused by the test instrument. This standardization seems to cause systematic errors (Hauenstein & Foti, 1989). It is likely that errors from a test method uncontrolled ruin a valid score interpretation as an indicator of language ability. A certain biased effect (systematic/random errors here) might affect the observable score (the change of the true score) depending on various kinds of measurement conditions (Goodenough, 1950). This argument looks plausible in that they did not provide a clear definition of quantified random errors and systematic errors with empirical evidence in the experimental situation. Random errors may occur for unpredictable reasons, and in terms of rater reliability, they may be related to a lack of internal consistency within a single rater's rating patterns. It should be clearly demonstrated what random errors are and what systematic errors so that they can be quantified with data, rather than just a conceptual discussion. For example, conceptually, random errors may affect the internal consistency of a rater. The sources of random errors are unknown and unpredictable reasons, such as fatigue, personal preference, or emotions.

However, the sources of systematic errors can be slightly easier to find than those of random errors. Biased patterns in rating could indicate a systematic error. Leniency or severity in rating patterns, halo effects, and contrasts/similarity effects could be considered quantified systematic errors. In practice, the boundary between random errors and systematic errors is not clear. Both systematic and random errors in rating look like statistical discrimination in ratings within and across examinees. It is likely that the sources of systematic errors might contribute to making an idiosyncratic rating pattern, and this pattern might also affect the sustainability of consistency as well as the random errors. McNamara (1996) supported this argument, claiming that the number of errors can be reduced by breaking them down. This indicates that errors, whether random or systematic, are estimated through the analysis of idiosyncratic rating patterns, in spite of the uncertainty of the error property. This kind of categorization for measurement errors may contribute to finding a solution to how we can handle issues of rater reliability. Moss (1994) suggested a standardization of test procedures to resolve measurement errors:

> Where acceptable levels have not been reached, recommendations for enhancing reliability without increasing the number of tasks or readers beyond cost-efficient levels have typically involved (a) increasing the specification of tasks or scoring procedures, thereby resulting in increased standardization…(p.6)

One possible way to resolve these controversies is by employing a standardized training program. Regarding issues of rater reliability, a major goal of rater training is to teach raters to acquire inter-and intra-reliability, and to reduce the random/systematic measurement errors. In performance tests, this classical role is highly valued in obtaining high rater reliability, because the role of the rater may be as (or even more) powerful as that of the test tasks in evaluating language performance.

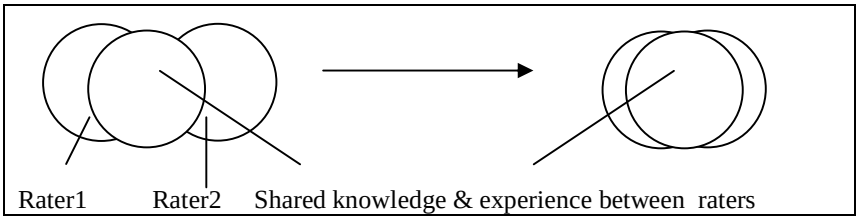


Figure 2. The process of agreement between raters.[1](Jang, 2006)

Reid (1993) presents an implication for rater training. As shown in Figure 2, the concept of inter-rater reliability can be reconstructed. During a training workshop, raters communicate with each other and with the trainer to increase shared knowledge and decrease ambiguity through the allocated rating tasks. Increasing shared knowledge and rating schema, therefore, helps raters comprehend rating tasks and facilitates the understanding of the intended test purpose or rating purpose. Such a standardized rater training program adjusts the balance between knowledge and experience of raters.

Recent research (Brown, 1995; Choi, 2000; 2002; McNamara, 1996; Saal, Downey, & Lahey, 1980; Shin, 2001; Steward, 1999; Weigle, 1994a; 1994b) has consistently shown that rater training plays a role in reducing ambiguity and maximizing shared knowledge among the raters, although there exist individual differences in rating schema, background knowledge, and experience in essay rating. Through a standardized training program, sufficient information on the rating process itself and the rating context are necessary for avoiding or reducing rater bias.

Educational impact of standardization. Standardization would be necessary to reach the standards of accountability, ethics, and fairness, as well as higher reliability in language testing. Accountability can be acquired not only at the test level (or program level), but at individual level as well.

[1] This figure was developed based on reading process model of Reid (1993) and adopted from Jang's early research (2006) .

At the program level, standardization would contribute to shared responsibility among test users. Since a standardized training program offers a chance to use the test instrument appropriately, test format effects could be reduced, and as a result, measurement errors could be eliminated (Mathison, 1992). Standardization of test procedure would present an alternative way to reduce the effect of test format.

At the individual level, accountability can be achieved in the form of professional improvement. It is expected that the professionalism of the raters could be improved by providing access to similar input. High quality information contributes to increased retention of knowledge and awareness through reinforcement of learning based on iterative feedback. Raters are made aware of the importance of their responsibility for scoring with respect to fairness. A deeper understanding of the rating context, such as the purpose of the test and the characteristics of the test takers through a training workshop could reduce a gap between raters' idiosyncratic expectations and the intention of the actual assessment instrument. These aspects of a standardized training program can contribute to enhancement of rating validity and scoring reliability, as well as to a positive washback effect of the program (Mathison, 1992).

Weiss' program evaluation theory supports this view in the sense that a combination of program theory with implementation theory redefines a program's theories of change (Weiss, 1998). Theory-based program evaluation plays a role in showing the entire framework: what kinds of activities are implemented to make changes, how the effectiveness of the rater training program can be evaluated, and what the expected program outcomes are. Program theory of change shows what kinds of program activities will take place and what will happen after the implementation of these activities.

Figure 3 shows how a standardized training program would positively influence a rating system through a chain of interim outcomes and the desired results for the ultimate program goals. The diagram devoted to the chains of desirable effects shows the relationships among program inputs, specific program activities, interim outcomes, and desired results in a different way. Educational materials (input), resources, and the rating environment for the training raters will be considered based on this relationship. In addition, how the program activities for training raters will proceed, as well a means to measure interim outcomes and results will be taken into account (Kiely & Rea-Dickins, 2005; Lynch, 2003; Weiss, 1998).

Based on Figure 3, the intended program goal is to enhance rater reliability and to provide a more supportive rating environment. One assumption is that a standardized training program is available (at all) to raters (it may not be). The first theoretical claim could be that if such training is available, then raters have an opportunity to understand and improve their rating skills. More structured materials and a better understanding of the skills would be helpful in improving their interview skills and thus their rating reliability. Another theory is that a standardized training program would guide a trainer's administration of the training a rater group. The trainer would encourage raters to carry out their tasks, giving them a higher level of motivation, and would help lead them to a greater self-awareness. A third theory may be that the raters and trainer could share the special problems raters encounter during the interview and rating sessions. The raters, trainer and staff of the testing agency cooperate to observe, become aware of raters' problems, and have a chance to find multiple ways to increase reliability. Figure 3 was developed based on Weiss's (1998) logic of the effectiveness of program evaluation.

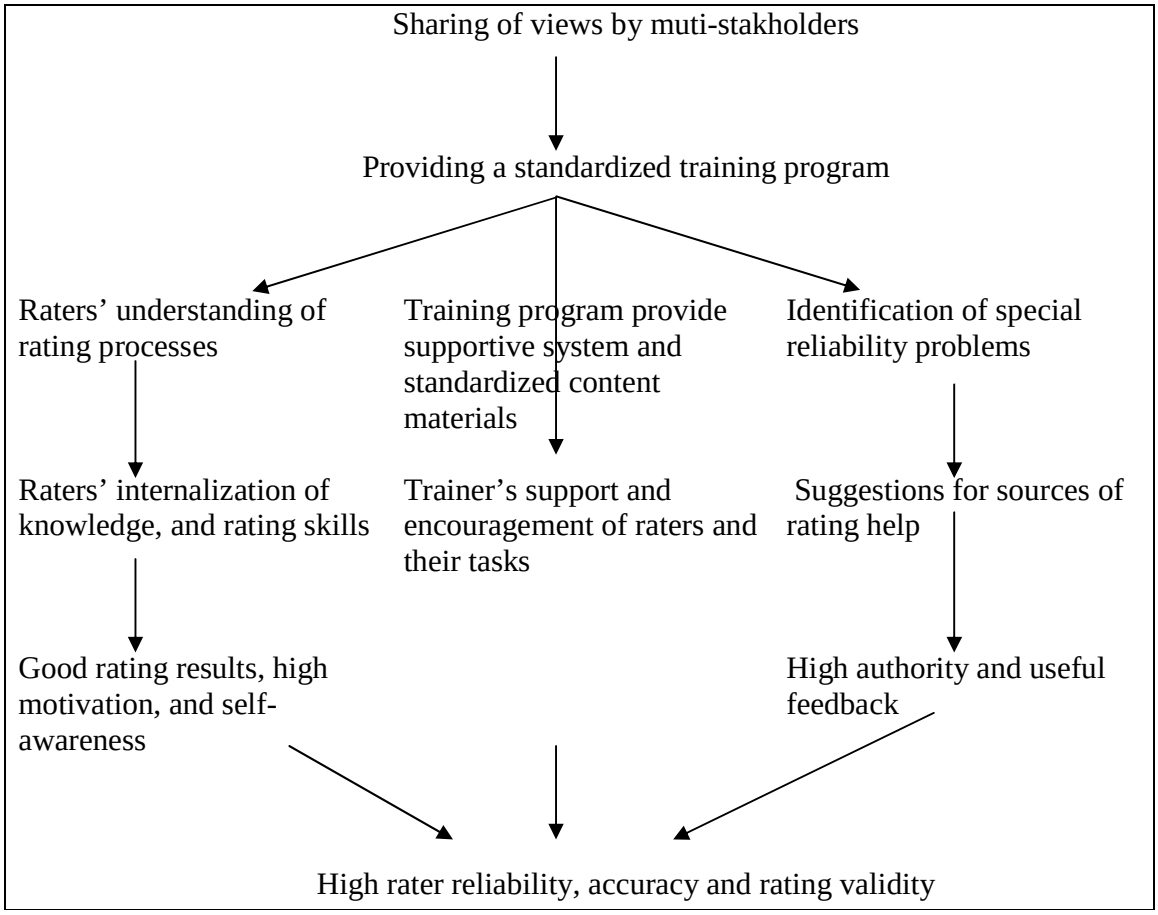


Figure 3. Program theory model for evaluation of a standardized training program.

Standardization would enhance the ethics of test use because it insures quality training information, and increases score fairness by improving the quality of individual raters' rating performances. Standardization of a training workshop is one way to establishing cumulative growth in rating (Lynch, 1996; Phillips, 1997; Rothwell & Kazanas, 2004; Waagen, 2006).

Features to be Considered for Standardization

Of primary concern in this chapter is what kind of factors should be considered to achieve standardization for training effectiveness. Three factors for standardization

should be considered: the rater group, the test itself, and the theoretical framework of the standardized training program. In general, training effectiveness reflects how well a standardized training program operationalize its theoretical test constructs in the workshop program, and how well the test procedures can be systematic to reflect the authenticity and interactiveness of the test tasks. The extent of rater reliability can be clearly defined in training workshop by analyzing raters' concerns. Therefore, this study suggests that all possible considerations which may influence the score reliability should be reflected in the rater training program by providing information about contextual factors within the rating environment, rather than focusing on a certain single aspect (Eckes, 2008; Lumley & McNamara, 1995; Lumley, 2005; McNamara, 1996; Shohamy et all., 1992).

First consideration: raters. When designing a training workshop, the primary consideration should be the raters, who are the consumers of the training workshop. DeRemer (1998) suggested that two aspects, knowledge and process, contribute to understanding the activity of rating. Here, "knowledge" refers to the general information relevant to the scoring procedures and the rating environment. "Process" refers to the raters' cognitive judgment process during scoring. Considerations about these two influences on rating performance should be understood and taken into account before conducting a rater training workshop.

Rater's cognitive process model in rating. Furneaux and Rignall (2007) asserted that understanding how raters reach their final decision in the rating process is one source of knowledge, because raters tend to consider multiple aspects rather than focusing on a single aspect when making judgments. To understand what is happening in the raters'

minds, we need to take a careful look at the hidden cognitive process of raters. The cognitive process is based on a combination of several different behaviors: reading, interpretation, and judgment (Borman, 1978; Cooper, 1981a; Freedman & Calfee, 1983; Lumley, 2005; Payne, Bettman, & Johnson, 1993).

Borman (1978) said that the cognitive or judgment process was defined as observation, evaluation, and weighting of the evaluations. Cooper (1981b) elaborated upon Borman's model in more detail: behavior observation, encoding, storage, retrieval, and evaluation. Freedman and Calfee (1983) presented the information-processing model of rating with variables affecting the rating process. This model is developed based on a three-step process: reading, evaluating and judgment. Lumley (2002) suggested three stages of scoring: reading, pre-rating, and confirming the assigned score.

In figure 4, the rating process was elaborated upon and reconstructed based on the Borman model, the Cooper model and the Freedman and Calfee model. This cognitive model includes four major steps: observation, selection, anchoring, and judgment. In the first stage, raters carefully observe the language corpus (essays or speech samples) by simply reading or listening to their rating assignments. On the basis of the information gathered and stored in their short term memory, raters try to select the relevant information by retrieval. In this second stage, raters have a chance to make initial assessments of their assignment. This is similar to individual interpretation of what they have read. The next stage is the anchoring stage, in which correct perceptional categorization and balanced weighting on a score may occur in the minds of the raters. Through the adjustment process, finally, they reach their own decision for a final score, after confirming their initial decision.

The salient feature of this model is the anchoring stage, between the selection and judgment stages. It is likely that rater variability might start to occur at this stage, based on how raters select and use the relevant information allowed. Payne et all. (1993) said, "Adjustment represents the transformation of the internal scale value into the external response scale." The process for anchoring demands more a complex cognitive process: retrieve and select information, integrate information, and form responses via their own problem-solving process (adjustment). Internal judgment is expressed on the external scale by iterative dynamic adjustment - in other words, a match between the anchor (the scale descriptors or assessment criteria) and the target value to be estimated (essays).

Homburg (1984) suggested conceptual "categorization" for scoring before moving on to evaluation of the essay in the rating model. Categorization is a significant method of correctly forming internal rating scale categories closely matched to the operational rating categories. Raters retrieve their memory stored in order to compare the individual anchoring point to the target value to be estimated, and pay attention to the target feature similar to the anchor (Chapman & Johnson, 2002), which particularly refers to a rating scale under rating context.

The categorization stage is meaningful for valuation and expression in judgment, but raters are asked to consider key characteristics at the anchoring stage between selection and evaluation, and these simultaneous requirements can burden a rater's cognitive process. It is likely that incorrect categorization in the initial anchoring process or a different information gathering process could result in a bias pattern in rating or interaction effects (Chapman & Johnson, 2002; Keysar & Barr, 2002). Kahneman and Frederick (2002) discussed weighting biases--which people would over-weight or under-

weight one salient feature--because a weighting bias would be unhelpful to the justification of a final decision by the rater. Gilbert (2002) called the adjustment process the "correction processes," and Wilson, Centerbar and Brekke (2002) explained it as a monitoring process that means "being aware of something." This means that self-monitoring for correct categorization at anchoring stage might be helpful to avoid a bias or mental contamination, which is an unwanted mental process.

The judgment process is not a simple process and should be distinguished from the notion of choice which is derived from personal preference (Payne et all., 1993). Judgment and choice have similar processes for selecting an alternative. The judgment process, however, is more complex in that an evaluation process is included. In other words, the judgment process demands more cognitive effort for raters to reach a reasonable decision. Therefore, raters often set up a model of the decision making steps in order to simplify the cognitive process, and raters sometimes employ various strategies, such as ignoring or eliminating less relevant information (see Figure 4).

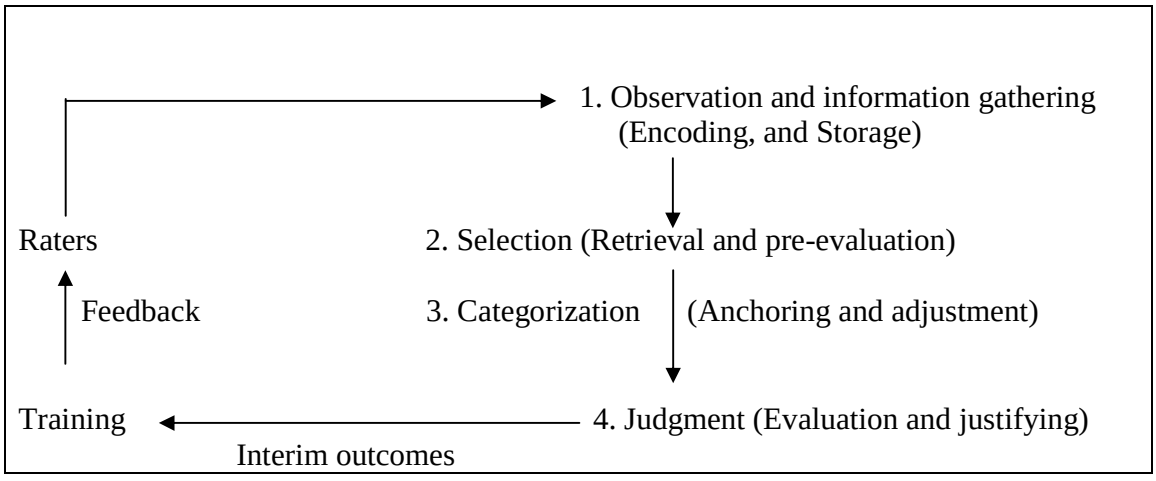


Figure 4. Rating processes.

Rating Strategies. Adaptive decision behavior refers to "flexibility" in decision-making (Payne et all., 1993). Decision makers not only tend to use only explicitly displayed information, but also might infer missing values. The use of adaptive decision strategy fails due to a deficit of decision knowledge, information display factor assessing tasks and contexts, and overreliance (or overconfidence in) an inappropriate strategy. Conflicts among values often create decision problems. Individual raters frequently employ multiple decision strategies in different situations to solve a conflict.

Based on the a review of the literature (Cummings, 1990 ; Erdosy, 2004; Lumely; 2005), Figure 5 summarizes the possible decision strategies for each stage from Figure 4. It is likely that different decision strategies (Payne et all., 1993) [2] could be employed depending on rating stage, the properties of the test task, and the characteristics of the language text to be evaluated. In the first two stages--observation and selection--simple strategies are employed, such as interpretation strategy and pre-assessing strategy. These strategies are used to understand the language text and scale descriptors (see Figure 5). Milanovic, Saville, and Shuhong (1996) suggested four different approaches that raters might take in the essay evaluation process: principled two scan/read, pragmatic two scan/read, read through, and provisional mark. The principled two scan/read approach seems to occur when raters try to give a final score after reading the essay twice. While reading, some raters may focus on specific features, while others pay attention to a general impression for rating. The pragmatic two-scan/read approach occurs when raters face some difficulties in deciding a final score. Raters read the essays several times to

[2]A drawback of these rules is that it is not very plausible that these strategies would perfectly apply to an actual judgment process, since this kind of strategy is for choice, rather than the judgment process. These decision strategies, however, can be modified and offer possible decision strategies raters could use in their decision making situations.

resolve their problems. In the read through approach, raters evaluate the essay by trying to see good features and bad features. Finally, in the provisional mark approach, raters may repeatedly read portions of the essay and pre-rate the essay before reading the entire essay. Raters try to confirm their initial rating of a particular essay by finding evidence to support their initial assessment.

The categorization stage looks more complex than the previous stages, since raters can employ a variety of strategies in order to address their concerns before moving on to a final decision[3]. For instance, the monitoring/managing strategy is used when raters reread the text, keep their own pace for managing rating time, and attempt to anchor the examinees' responses to the rating scale categories. Simplifying or refining strategies may be adopted when raters are faced with more complex problems. Raters ignore the irrelevant information, or they reduce the amount of information processed. Raters sometimes attempt to process all the relevant information (Tversky & Kahneman, 2002), or they attempt to elaborate or redefine their information from both the scale and the language text by adding more information through a gap-filling process. As a

[3] Payne, Bettman, and Johnson (1993) have introduced seven strategies to be used in decision making: the weighted additive (WADD), the equal weight (EQW), the satisficing heuristic (SAT), the lexicographic heuristic (LEX), the elimination-by-aspects (EBA), the majority of confirming dimensions heuristic (MCD), and the frequency of good and bad features (FRQ). First, the weighted additive rule states that a decision maker considers all the relative importance or weights of the attributes. The conflict among values is resolved by the relative importance or weights, and an alternative with the highest overall evaluation is chosen by adding all weighted attribute values. Second, the equal weight is a special case of the WADD and considers all the alternatives and the entire attribute values for each alternative. It simplifies the decision-making process by ignoring the relative importance in order to make a decision accurately in a risky situation. The alternative with the highest average payoff is selected. Third, the satisfying heuristic compares the values of each attribute of an alternative to a predefined cutoff level. Next, in the case of the lexicographic heuristic rule, the alternative with the best value on the most important attribute is selected. If two alternatives have tied values, the second most important attribute is considered until the tie is broken. Next, elimination-by-aspects states that the alternatives with the attribute are chosen by rejection or eliminating alternatives below the cutoff. The majority of confirming dimensions heuristic is a kind of the pair-wise comparison, comparing the values of the two alternatives and selecting the better value. Finally, the frequency of good and bad features is a decision rule in which alternatives are selected based on counts of the good or bad features of the alternative.

comparison strategy, quantitative reasoning concerns the degree of the quantities via counting frequency, on the other hands, qualitative reasoning is used simply by comparing values (Payne et all., 1993).

The compensation strategy can be employed when raters consider subsidiary factors, such as the difficulty of test items or the examinee's proficiency level. The importance of weight may be different depending on the degree of correlation between the anchored rating scales and the language corpus assigning in rating. Raters can selectively take information or consider all information, so that the rating processing is holistic or dimensional (Tversky & Kahneman, 2002). When raters encounter a questionable score, they are likely to adopt a compromising strategy by rejecting or accepting scale descriptors, although they do not fit well. At the final step, raters attempt to confirm their final decision by using a justifying strategy. Raters justify their final scores based on factors beyond the text and rating scale descriptors. They consider a balanced score distribution or score pattern, score fairness, and overall quality of the text.

Stage	Decision Strategy	Focus	Behaviors
Observation	Interpretation strategy	Text	• read language corpus • identify overall organization • summarize ideas • interpret ambiguous or unclear part • edit phrases for interpretation
		Task	• read/understand test task/prompt • interpret task requirements
		Scale	• read/interpret the scale category
Selection	Pre-rating strategy	Text	• pre-rate reasoning, and logic • pre-rate topic development • pre-rate coherence/organization • identify redundancies • identify error frequency/classify the types of errors

Figure 5. Raters' decision strategies for each stage.

Figure 5 (Continued)

			• pre-rate level of vocabulary • pre-rate syntax and morphology • pre-rate spelling and punctuation
		Task	• pre-rate task completion • pre-rate relevance to the topic given
		Scale	• defining the scale category
Categorization	Monitoring/ managing strategy	Self-Monitoring	• assessing general impression • reread language corpus
		Scale	• distinguishing the scope of scale categories by expanding/ narrowing the scope • managing conflicts within/between scale categories
	Comparison strategy	Task	• compare tasks within a test
		Text	• compare with other test takers
	Compensation strategy	Task	• compensate across tasks by considering the level of difficulty(or challenge) test prompt given
		Test takers	• considering test takers' background/ language learning • considering writer's understanding of prompt
		Scale	• weighting on particular traits or assessment criteria
	Compromising strategy	Scale	• accepting a score that is quite not match with the scale • rejecting the scale when the scale descriptors are unhelpful or unfair.
Judgment	Justifying strategy	Others	• elaborate the scale descriptors which did not explicitly show in the scale category • incorporate features absent from the scale descriptors • consider balance of scores within a text • judging text overall quality, insufficient text, questioning scores

To sum up, individual raters could score differently to a wide variety of task conditions and task environments. Understanding the cognitive process and the use of decision strategy can provide implications for preventing variability from different

cognitive processes, cognitive distortions coming from short-term memory problems, incorrect categorization, different problem solving processes and carelessness and for designing the training program around these factors. There is a valuable implication for rater training: in terms of cost effectiveness of thought, the cognitive processing should be simplified, and appropriate decision strategies should be identified and standardized. It is expected that these efforts could enhance scoring reliability.

Second Consideration: Rating Environment Instruments.

Sources of bias in the rating environment. While scoring, a rater is asked to multi-task, so that the rating performance is influenced by the complexity of the rating task, including factors such as the examinee's response, rating scale descriptors and assessment criteria, the topic or task constraints, and time pressure. Individual raters respond differently to these multiple sources or rating constraints. DeRemer (1998) found that raters had different rating focuses: a global, impression-based rating focus; text-based judgment; and scale descriptor-based judgment. This solution process causes some degree of interaction or bias effect, such as a rater's bias toward an examinee group, assessment criteria, or test task. The degree of these effects may determine a rater's consistency or rating accuracy.

In addition, some researchers have found that individual raters' subjective judgment causes rater variability (or error) because of the complexity of their individual background, such as rating experience, language and teaching background factors, and training experience (Brown, 1995; Choi, 2000; McNamara, 1996; Saal et all., 1980; Shin, 2001; Steward, 1999). It is well-known that rater groups with different backgrounds (native or non-native speakers, having field-specific experience or not, etc.) have shown

differing attitudes towards the language corpus itself and thus approach or interpret rating situations differently (Shoham et all., 1992). This section will review the arguments with respect to the possibility of interaction effects between raters and any facet which is involved in scoring environment. Since these interaction effects are consider as a systematic rating pattern which might be reducible, this effort is significant to identify the direction a critical design element of any standardized training workshop (Schaefer, 2008; Wiggleworth, 1993).

Different perceptions of examinees. Kondo-Brown (2002) and Schaefer (2008) investigated the interaction between a rater and examinees. In Kondo-Brown's study, it appeared that raters showed some bias depending upon the proficiency level of the examinees. Raters had bias patterns toward examinees who placed at extremely high or low proficiency levels. They tended to be more severe toward examinees at extremely high proficiency levels, while they had a tendency to be more lenient toward examinees at extremely low proficiency levels. These findings suggested the establishment of clearer standards regarding the interaction between raters and specific aspects of the examinee.

Rating issues such as interaction effects between particular rater and examinee groups may arise if raters do not fully understand the audience group being scored. It was found that raters have different perceptions of the language development stage, the definition of the best writing in an ESL setting, and good or bad features of ESL writing. This indicates that individual raters have different expectations of the best response, so that discussion about these features could enhance score reliability.

In addition, raters also consider factors beyond the response to the test prompts, such as score fairness. With regard to the intended test taker audience, more specific

information on the characteristics of the group should be provided in order to facilitate the rating process. For example, proficiency language test has broad audience layers, meaning that the examinees' proficiency levels vary widely from extremely low to extremely high. Many studies have found that raters have some bias on the extreme ends of proficiency levels.

On the other hand, when the test-taker group is likely to be more homogenous, such as when a college board determines assistantship hiring on the basis of test results, evaluations tend to have different sets of standards. In this case, since candidates for assistantships have already reached a certain similar level of language proficiency, the candidates tend to be evaluated on a relative scale based on the performances of the others who take the test.

Another issue of some relevance here is the matter of score use. It is fairly difficult to determine how scores or results will be appropriately used by stakeholders. So the issue of score use is an important factor to be considered in designing a training program and should be shared with raters, since each rater's decision process benefits from knowledge of the scoring objectives. Test results may or may not have powerful washback on an individual's progress, depending on how a score is to be used. In addition, a rater's decision might be different depending on how the rating result is to be used.

For example, a score can be used as a cut-off point for passing or failing in education programs such as ESL courses (low stakes) or used to determine admission to a graduate course (high stakes). If a score is used for a simple pass or a fail, raters focus on giving helpful feedback to test-takers to help them improve, but if a score is used in a

high-stakes decision making process, raters tend to be more strict in their scoring This, then, should be a consideration that should be reflected in rater training programs. Information about the examinee group is expected to help raters adjust how leniently or severely they score the response given.

Certain effects may surface, such as contrast effect, halo effect, central tendency in scoring patterns, and similarity effects between raters and examinees. The order of the essay/speech sample rating might cause a contrast/comparison effect. For example, if a rater scores a highly proficient language performance first, and then comes across a speech or essay with a lower proficiency level, the rater might feel a bigger proficiency gap exists between the two examinees. Consequently, the second rating might be lowered because of the biased impression caused by the higher-proficiency first examinee.

Different perceptions of test tasks/prompts. Hamp-Lyons and Mathias (1994) investigated the difficulty of different types of test prompts. In an ESL situation, private and expository prompts are considered more difficult than public and argumentative prompts. These findings were contrary to the general assumption about prompt difficulty, but they make sense, since ESL students have more opportunity to practice responding to public and argumentative prompts in their learning. Schaefer (2008) discusses to the difficulty of ignoring task effects on rating. Raters may sometimes adopt compensatory strategies for examines who have a more difficult test prompt or test task. Weigle's study (1994a) demonstrated that both experienced and inexperienced raters showed some bias toward task type. Inexperienced raters were more severe on graph tasks, but experienced raters scored more strictly on choice tasks, because they assumed that choice tasks are easier task than graph tasks. These findings suggest that the effect of test tasks/prompts

on rating performance still exist, and some discussion is needed to reduce interaction effects.

Raters are sensitive to the difficulty level of the task, depending on the type of test prompt (Shaw & Weir, 2007). Also, they have their own expectations about what the best answer looks like. When multi-test items are used, raters may use compensation strategies, depending on the type of task prompts or the difficulty level. Raters tend to give more lenient scores on more difficult tasks, and conversely, they tend to score more severely on easier tasks. It is likely that raters strongly consider score fairness and give more benefit to the examinees who encountered more difficult tasks. Guidance on how to manage this situation needs to be discussed and provided via consensus among group members.

Different perceptions of assessment criteria. Raters have their own rating patterns of scoring severely or leniently on a particular criterion. In particular, based on a review of the literature, raters have different rating focuses (key features), different weighting on criteria, and different conceptual hierarchies. Kahneman and Frederick (2002) suggested that a lack of objective criteria and insufficient information are considered sources of bias in the decision making process. Some guidelines should be discussed and suggested in the workshop; for instance, key features in the operational rating context should be defined, and principles should be provided about how to narrow down their rating decision. Conceptual hierarchies among assessment criteria should be reduced, and, at the same time, assessment criteria should be weighted as equally as possible.

Numerous studies of interaction between raters and assessment criteria have been conducted (Eckes, 2008; Schaefer, 2008). Different weighting of assessment criteria is one factor that affects rating performance. Eckes (2008) has studied different rating styles by analyzing rating patterns using two-mode clustering. His research focused on identifying the relationship between 64 well trained raters and nine assessment criteria used in the writing tests. Raters were asked to evaluate the importance of each assessment criterion using a four scale band. Six rater categories were found and depicted in a hierarchical conceptual map. Cumming's study (1990) presented different results in analyzing the rating patterns of both inexperienced and experienced rater groups. The experienced rater group had a tendency to more equally weight the assessment criteria in comparison with the inexperienced group.

Moreover, rater background, including teaching and rating experience, might influence the construction of different weighting or identification of key features in rating. Many raters are language teachers who are native speakers of the target language. Teaching experience has a tendency to interact with assessment criteria, so depending on their teaching goals and their previous education program they have received and emphases, it seems that teachers rely heavily on specific components for their rating decision. Brown (1995) investigated the effect of job experience by analyzing the performance of three groups of raters for the Japanese test for tour guides. Brown's study of raters' different performances depending on background provides us with an interesting discussion. In her study, the major source of rating differences was job background, rather than language background. Her findings are not surprising, because generally teacher-raters consider linguistic features to be important in language classes.

Moreover, teacher-raters are likely to be more tolerant of student mistakes and poor performance in communication. Industry raters, however, assigned harsher scores in these components because industry-raters regarded the ability to deal with clients as a crucial factor in real situations.

In addition, teacher raters produced overfit rating patterns which do not clearly distinguish among scale levels because they used narrow range scales when rating. This result corresponds with other research (Bachman, Lynch & Mason, 1995; Brown, 1995; Shin, 2001) showing that most teachers are reluctant to assign either extremely high or extremely low scores to their students. It is highly possible that raters with tour guide experience are affected by non-language factors like body language and facial expressions, since raters assigned scores to a videotaped interview. It is obvious that industry raters showed some bias on specific tasks and components they prefer. These findings suggest a need to discuss how to adjust different weighting of criteria, and how to identify rating focus and key features used in the rating process.

Use of rating scale and its descriptors. Bernardin and Walter (1977) showed that rating quality might be influenced not only by the training format but also by the quality of rating scales. Raters' preferences about scale types had to do with different rating performances. Some preferred a simple scale type, but others preferred a more sophisticated type. Cooper (1983) pointed out that discrepancies between descriptors and students' performance and between descriptors and the purpose of a test might have a negative influence on raters. Lumely (2002) pointed out the matter of irrelevant and unclear descriptors. His findings pointed to the need to consider what kinds of rating scale descriptors help raters maintain more consistency in their rating.

Rating scales and descriptors show theoretical directions such as examinees' language competence on a more abstract level, but they should also reflect the actual performances of examinees—their ability to use language in lecture and practice sessions. Lumely (2002) pointed out wording problems of scale descriptors in terms of relevance and clarity, and many researchers have emphasized the development of descriptors based on empirical evidence in order to reduce this problem. Fulcher (1987) insisted on the development of criteria based on empirical evidence via direct observation of the students' performances. For example, "can-do statements" have been used to explain proficiency levels.

Standardized rating scale descriptors would be one premise in reducing error in rating through empirical study. Several researchers have shown the development process of descriptors via discussion of rater groups (Fulcher, 1987; Upshur & Turner, 1995; Weigle, 1994a; 1994b). Weigle's findings have implications for a well-designed training program, suggesting that ineffective descriptors and scales should be readjusted or revised by discussion among raters. Upshur and Turner (1995) also showed how to develop valid scale descriptors and introduced the applicability of a binary rating scale method for improving rating performance. Alderson (1991) suggested the development of an assessor-oriented scale that describes the observed performances of students in order to minimize the influence of scales.

Raters are capable of distinguishing what a test taker within a certain level can or cannot do in order to make an accurate decision. To achieving correct categorization, Cooper (1983) has suggested that crafting more solid scale descriptors is necessary to keep away from incorrect categorization. It is impossible for scale descriptors to describe

all possible behaviors of examinees at each proficiency level (Sulsky & Day, 1992), but in their study, more specific scale descriptors and clear rating categories tended to reduce rating error. It is suggested that the scale descriptors should include the prototypical features of proficiency levels, which is including consistent features with examinees' behaviors (Cooper, 1983; Slusher & Anderson, 1987). It is expected that empirical scale development might aid raters' decision making processes in terms of what kinds of aspects of an examinee should be observed, and what kind of information a rater should select to avoid information loss.

Scoring methods. Brown (1995) and Weigle (1994a) also found that native raters depended considerably on their intuition, although a discrepancy that native speaker raters were harsher than non-native speaker raters appeared in Brown's study. Cumming (1990) found distinguishable behavior patterns: the experienced group tended not only to pay more attention to the quality control aspect at the abstract level but also to have a more holistic rating focus. Moreover, they also had clearer focuses or categories in rating and their own strategies for controlling rating problems they had met. Erdosy (2004) supported Cumming's study in that more experienced raters not only preferred global impressions, but handled rating successfully with a simple rating process despite the consideration of multiple factors involved in scoring the assignment (DeRemer, 1998; Homburg, 1984).

Cumming (1990) suggested that an analytic scoring method may be disadvantageous to examinees at lower proficiency levels. The separate guidelines may be ineffectively used to distinguish different aspect of essays within lower proficiency levels. In spite of this disadvantage, for the purpose of rater training, analytic scoring

methods are advantageous to inexperienced rater groups, since they seem to display a narrowed rating focus. Shi (2001) has discussed the agenda of holistic versus analytic methods. Shi stated that the holistic approach has the disadvantage of failing to discern accurate proficiency levels. The analytic scoring method looks more effective in this situation, because it can provide a higher quality of information about the examinees.

Discussions about scoring methods are also important. It is still inconclusive which methods--holistic or analytics scoring methods--improve rater reliability (Shin, 2001; Shin & Jang, 2002; Varghan, 1993). The use of both scoring methods is suggested during the practice session. An analytic scoring method is recommended at the beginning of the workshop, because it may be more helpful for less experienced raters. After acquiring a minimum level of rating experience, a holistic scoring method is recommended, which reflects the behavior of experienced raters.

Third consideration: Theoretical Framework of a Training Program. A poorly planned training program without theoretical models does little to enhance rater reliability. For instance, sometimes a module should be clearly specified, and the roles of the entire stages from pre-workshop to post-workshop stages should be well-defined. In the next section, past rater training models will be introduced.

Past training models. The design of a training program determines the basic structure of the training program--what the main focus is, and what kind of training the assessors should receive. Woehr and Huffcutt (1994) historically reviewed the characteristics of rater training programs. A general framework of rater training programs categorizes them into four types, according to the training content or focus in determining the effectiveness of the training appraisal: rater error training, performance dimension

training, behavioral observation training, and frame-of-reference training. Meanwhile, one problem with these four different training models is that most training programs are related to rating issues, and therefore, consideration of interview procedures, any other test instruments, and any modifications is necessary in order to apply these four theories to the concept of standardization.

Rating error focused training program. A psychometric approach is an evaluation method in which rating errors are found and their sources are identified in order to achieve improvement. The earliest approach to the evaluation of rater training programs was rater error training, emerging from the psychometrician's view of performance ratings. In the past, most rater training programs have made efforts to estimate rating errors and suggest how these errors could be estimated with a quantitative approach (Cooper, 1983; Engelhard, 1994; Latham, Wexley & Pursell, 1975; Levine & Buttler, 1952; Saal et all., 1980; Shin & Jang, 2002; Spool, 1978). When designing current rater training programs, reducing rating errors is neither a primary goal nor an index of the effectiveness of a particular training program, although a reduction in rating errors might enhance rating accuracy (Smith, 1986; Spool, 1978). In terms of issues of reliability, the focus of a training program is how consistent raters are in their decision making. In order words, these rating errors could be considered idiosyncratic characteristics of raters rather than easily removable features (MacNamara, 1996). However, it is highly possible that these errors should be reconsidered as a major source of bias or interaction effects hampering reliability.

Some efforts to precisely measure rating errors, such as having a restricted score range, halo effect, contrast effect, and leniency seem to be made at a minimal level, and

in particular, these considerations are necessary during the training session. Rating errors are estimated by observing and analyzing rating patterns or the distributions of individuals so that they could identify the major source of the errors. In this study, the estimation of rating errors is one way to determine the effectiveness of a training program, since diagnostic information can be provided by analyzing the amount of variance for a rating aspect, as well as the consistency of rating patterns.

Performance dimension oriented training. According to Woehr and Huffcutt (1994), a second type of rater training involves the dimensions of performance in view of cognitive information processing with respect to rating accuracy. Raters are trained to identify and employ the appropriate dimensions, and this approach is partially adopted in the current training program in order to (re)train raters. Rating is improved by familiarizing raters with the dimensions of prototype samples with official ratings during rating practice. In general, this training program has focused on enhancing rating accuracy. Rating accuracy is computed by comparing individual ratings with ratings provided by an expert rater.

Bernardin & Buckley (1981) evaluated performance dimension training. Performance accuracy is a comparison between the individual performance dimension and an expert rating (true score) in terms of analytic scores for assessment criteria, and holistic scores. Through this activity, assessors become aware of the gap between their ratings and those of expert scores. The biggest advantage of this approach is that rating validity can be inferred from the accuracy of the two rating scores as well as reliability in rating (Sulsky & Balzer, 1988).

Behavioral Observation Training. Behavioral observation training focuses on observation of behaviors, which includes the detection, perception, and recall or recognition of specific behavioral events. Bernardin and Buckey (1981) discussed the fact that the observation process during the gathering of evidence refers to an ability to recall information more accurately. Behavioral observation measures the degree of recall of specific information, and rating errors interfere with memory recall. In their article, they investigated how much a rater correctly observed on video materials using questionnaires, and they compared three groups. The results suggested that rater accuracy in the group where the training lecture session emphasized observation accuracy was enhanced and rating errors were reduced. Thornton and Zorich (1980) classified some sources of error:

> Loss of detail through simplification, overdependence on a single source, middle message loss, categorization error (forcing observation into categories instead of remembering the differences between ideas, behavior, and people), contamination from prior information, contextual errors, prejudice and stereotyping, and halo effect (being overly influenced by one characteristics of a person). (p.353)

On the basis of the classification of Thornton and Zorich, frequent rating variability could be more easily explained. Information loss, categorization error, and rater's bias could be barriers to reducing scoring reliability. Another task is for raters to match observation of the multidimensional behaviors with rating categories. In addition, rater bias can be formed by past rating experience, preexisting stereotypes, and contamination from prior information.

The implications can be drawn that there are some possibilities for reducing rating errors via accurate observation of the behaviors of examinees. Accurate observation would prevent rating errors, such as halo caused by information loss through over-simplification and assimilation to prior information. One shortcoming of behavioral

observation training is that this approach focuses only on behavior observation, rather than rating activities using both observation and evaluation of behavior (Thornton & Zorich, 1980). However, this type of training has some impact on increasing inter-rater agreement (Spool, 1978).

Frame -of -reference training. The fourth type of rater training program is the frame-of-reference training program, which considers simultaneously multiple dimensions of performance and performance standards. Frame-of-reference training encourages raters to share and use common conceptualizations of performance when scoring a language corpus. Current training programs look like frame-of-reference training because they share the general standards of individual testing agencies. The primary role of the frame-of-reference training program is to introduce raters to the entire rating procedure, step by step. The major premise of this training strategy is that raters are trained with representative samples, practice, and feedback, based on the standards of the level of performance being assessed. The quality of rating can be easily enhanced through standardized training because participants are given standardized information and share a common framework. This approach is distinguishable from the previous three approaches because it emphasizes several points such as general standards for ideal rating, valid descriptor development based on empirical study of descriptors, and examinees' responses.

Training focus: consistency, level of severity and accuracy in scoring. Many studies (McNamara, 1996; Furneaux & Rignall, 2007; O'Sullivan & Rignall, 2007; Shaw, 2002; Weigle, 1996; Wiggleworth, 1993) have shown how much rater training affects rater behaviors in terms of scoring reliability. It is likely that rater training influences

improvement of self-consistency. Raters tended to score more consistently after training. However, severity level (agreement) and accuracy, which can be estimated with standard deviation, were still arguable. In terms of severity level, raters tend to be relatively more lenient or more severe after the training rather than perfectly reducing the differences in their severity level (Shaw, 2002; Weigle, 1998; Shaw & Weir, 2007). Further research on the effects of training on increasing rating accuracy and agreement among raters is needed.

Current operational training programs. Jang's special qualifying exam paper (2007) reviewed the four currently operationalized training programs based on a document analysis of its training workshop book. The training program is divided into two categories: Face to face (off-line) training, and on-line training. In addition, each of these categories is divided into two types of test: speaking and writing. The American Council on the Teaching of Foreign Languages Oral Proficiency Interview (hereafter ACTFL OPI), Berlitz Proficiency Interview (BPI), and Multimedia Assisted Test of English (MATE) are rater training programs for speaking tests; the UIUC ESL Proficiency Test (EPT) is for writing tests. In this study, four training programs were chosen based on the review of relevant materials, both in hardcopy and online. Although three of the training programs in this study are related to speaking tests, it is likely that test format is not a primary matter, and the training workshop seems to have similar procedures for training raters across test types. Here, three face-to-face (i.e. off-line) training programs discussed and evaluated based on an analysis of the training materials (see Appendix A and B). The online training programs, however, were not discussed, because of confidentiality issues and limited access to online training materials.

One role of training is to bridge the gap between rater expectations and the actual assessment results, but rater expectations, the standards for evaluation, and a communication channel for two-way feedback are not clearly reflected in the training procedures. Based on a document analysis of the various training guidelines, rater training looks to be comprised of simple guidelines for introducing the structure of the tests, test tasks, assessment criteria, scales, and scoring procedures via a one-way delivery system. It indicates that the entire process for screening a qualified rater seems to lack a systematic procedure across the different tests under review, and the role of the training modules is unclear at each stage. It is highly possible that many training programs do not consider this matter, and the programs are arbitrarily designed, depending on institutional context or trainers' decisions.

As a balance between two major aspects of training--familiarization and norming—should be achieved when designing a program. In most cases, the focus of the training is on familiarization, in which general information about the test is disseminated. Based on this document analysis, three training programs seemed to emphasize the familiarization process over norming. Only one program included norming session. The norming session, conversely, should be more intensified. Further evidence of this argument is the lecture style of the training programs. A large portion of the training session is assigned to conveying general information and principles, across the programs. However, to achieve high rater reliability, more activities for norming—aligning raters' standards with the standards of the tests—should be developed for the training session.

Measures of rating skills have been calculated based on individual reliability estimates. In the review of the training guidelines, no reason was found to consider this matter of

measure. Particularly, in the language testing area, the major focus of the research has been intra-rater reliability or comparisons of raters' leniency or severity during training.

Three suggestions arose from this exploration. First, better clarity of the theoretical model and the well-organized content of the training materials are suggested for designing a more systematic training program. Theoretical models would prove guidance and clear direction reflecting both the nature of the test and an idiosyncratic rating context for the design of a particular training program. The theoretical models would cover the pre-training stage to the final feedback stage, and provide specified tasks for moving through the stages, as well as familiarizing, norming, evaluation, and problem-solving to enhance reliability.

The second suggestion for the improvement of training programs was that the training program provides general standards/principles for the generalization of the training information or materials. An analysis of speech/writing samples to refine scale descriptors both qualitatively and quantitatively should be provided, and in terms of the cognitive decision-making approach, an analysis of the assessors' performances should be conducted to understand their challenges in certain aspects of the rating and to identify their concerns in multiple ways, both quantitative and qualitative.

Finally, a training program provides the acceptable extent of scoring variations among raters. Proper scoring policy, as well as the administrative expectations of each training program, should be made available to the raters in the training materials. If possible, explicit statements and clear guidelines should be provided regarding what raters should and should not do. The training materials should reflect the entire screening process from applying for the workshop to becoming a certified assessor.

These three requirements can be satisfied by standardizing the training materials, training methods, and internal policy provided by the testing agency. Each of these approaches for establishing cumulative growth in rating is a different way of determining rater quality (Lynch, 1996; Phillips, 1997; Rothwell & Kazanas, 2004; Waagen, 2006).

Ways to Standardize a Training Program: New Directions for Standardization

This section gives a picture of the theoretical background of standardization to realize operational details of how standardization can be achieved in practice. The role of a rater training program and training guidelines, generally, would be to connect theoretical constructs of language proficiency to operational constructs of language proficiency. Training materials and methods reflect the test specifications proposed by Davidson and Lynch (2002). Test specifications deliver the theoretical constructs for language ability that a particular test pursues. Theoretical constructs can be ascertained from the test tasks given and the scoring procedures, and they are described through the assessment criteria and rating scale descriptors.

Overall discussions of the rating context. Fulcher and Davidson (2007) discussed specification-driven language test development, and "quality management systems" through which specifications can be applied to the test's operational stage and a particular test may be maintained by iterative feedback of. For instance, the training workshop is the place to accurately convey the intentions of the test developers, which indicate the theoretical definition of the constructs and the test task specification.

In line with the congruence procedure, efforts to fit the spec should be made during the rater training process to increase construct validity and reliability. Rater

training programs enhance the practical use of the test instrument by exhibiting the purposes of the test, the characteristics of the test tasks and scoring procedures, the characteristics of the target examinees, and test score interpretation. Training program/materials should be crafted with standards from broad vision statements to precisely indicate what raters are expected to know, what they should be able to do, and in what contexts they are expected to achieve the primary training goals. In addition, these expectations and standards should be clearly represented in the training materials. This implies that the operation of a well-designed training program leads raters to improve their rating performance and contributes to test validity by ensuring the appropriateness of use of the test instruments (Henning, 1987).

Formulating standards for an expert rater. Setting up standards for identifying a qualified rater is necessary to the training workshop. This standard provides not only valuable principles for rating but a direction for the program design of the workshop. These kinds of standards could be found and be modeled by analyzing the behavior of experienced raters in previous studies (Cumming, 1990; Lumley, 2002; Weigle, 1998). These studies have reported differences in rating behavior by comparing the rating patterns of experienced raters with those of inexperienced raters.

Salient features of experience raters are that first, they have structured reading steps and a specified rating protocol they usually follow, although rating variability exists among raters. Experienced raters can finish their rating tasks within a relatively short time. This indicates that through practice, their reading and rating processes have become internalized, and they are familiar with self-monitoring their evaluation behavior. Based on numerous reviews of past research, one interesting finding is that the cognitive process

of experienced raters looks simpler than that of inexperienced raters, because experienced raters tend to take a holistic approach. However, there are no patterns in determining rating focus among highly experienced raters (Cumming, 1990; DeRemer, 1998; Erdosy, 2004; Homburg, 1984). Kahneman and Frederick (2002) suggested two cognitive systems to express the cognitive operation. One is an intuition-based system, and the other is a reflective system. The reflective process seems more helpful for the production of evidence based reasoning. Shin (2001) investigated the differences between inexperienced and experienced raters in rating, and the differences between native speakers and non- native speakers of English with a combination of quantitative and qualitative methodologies. He reported that lack of rating experience might be considered as a potential source of error. In spite of adjusting the norming session in accordance with FACETS results, inexperienced raters presented lower self-consistency with misfit and were also relatively harsher than the experienced raters. He indicated that novices might have difficulty with both understanding the scales and with applying the scale descriptions to actual essay samples.

Secondly, experienced raters have internalized standards for rating, which help them sort and judge the language corpus. They tend to notice the quality of the language corpus and the types of errors, rather than surface features or the simple frequency of a particular aspect. Next, experienced raters maintain a balanced decision making process in weighting the assessment criteria. They are already aware of the key features in assessing discrete language abilities. They also adopt absolute standard as to whether a particular examinee reaches a certain proficiency level based on scale categories.

However, a hierarchical structure among key criteria is frequently found in the perception of novice raters. In addition, experienced raters seem more confident on their decisions, and tend to react positively to a violation of their expectations. They are likely to be more generous and understand the situation of the examinees, although they have met different approaches to the response. Conversely, novice raters tend to react more negatively on unexpected responses. Finally, it is likely that the decision making process of experienced raters is based somewhat on global impressions, but also on the collection of multiple pieces of evidence from the rating environment. Novice raters tend to score more severely or show extreme distribution in severity level (too severe or too lenient).

In spite of these common features of experienced raters extracted from previous studies, some agenda such as a unified rating model, more standardized rating focus, and similar definitions of language proficiency and language learning should be discussed to reach a consensus among raters during the training workshop.

Standardization of Rater Training Systems.

Standardization of developmental stages of a training workshop. The characteristics of a standardized training model could be suggested based on Jang's study (2006) and review of the second qualifying paper (2007). First of all, a standardized training program could be designed in three different stages: pre-workshop planning, monitoring, and post-workshop assessment; and with three different considerations: the appropriateness of the content materials and training methods, the interactiveness of the workshop, and the regularity of the program.

Figure 6 is based on the Bachman and Palmer (1996) model for the development of new language tests. This figure exhibits three stages of a standardized training

program: planning, monitoring, and assessing. In the planning stage, analyzing the characteristics of the test format and the stakeholders' needs is necessary. Depending on the results of the needs analysis, the focus/direction of the training content and the materials development workshop might be determined by a consensus among the stakeholders.

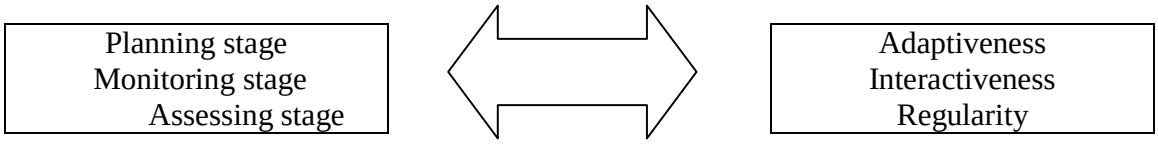


Figure 6. The three developmental stages for raters' training. (Jang, 2006)

Secondly, the monitoring stage, which occurs during the training workshop, is an opportunity for trainees to share their concerns discusses them to reach a group consensus. For instance, a balance between two major aspects of training--familiarization and norming—should be achieved by learning and practicing. In most cases, the focus of the training is on familiarization, in which general information about the test is disseminated. A large portion of the training session is assigned to conveying general information and principles, across the program.

To achieve high rater reliability, more activities for norming—aligning raters' standards with the standards of the tests—should be developed for the training session. A well-structured norming process bridges the gap between knowledge and application, which improves rater reliability. Weigle (1994a) investigated why new raters were more severe and less consistent than well-trained raters. Analysis of verbal protocol showed that inexperienced raters have problems when applying scale descriptors to essays because they did not participate in a norming session. It is common that individual raters might have some difficulty applying theoretical constructs to the scoring interpretation

process (operational constructs or the actual examinees' performances being rated) adjusting their inner rating scale with their outer rating scale.

Finally, at the assessing stage, the focus is on evaluating how appropriate and useful the training workshop was, including training materials, methods, and procedures. For improvement of rater reliability, the training materials/program should be strengthened by incorporating checking procedures, using performance observations, more accurate quantitative measures and deep analysis with qualitative approaches. Evaluation criteria measuring individual raters' improvement can also help estimate how well the standardized training program is functioning. Obtaining feedback from primary stakeholder on tasks and scaling descriptors that are judged as being problematic is important at this stage.

This approach can be used to verify and control the quality of the rating. Rater reliability can be determined by careful design of training workshop—how closely the rater expectations and assessments match—by interacting with individual raters.

The three properties of a training workshop. Based on the literature review, a systematic training program should have three properties: adaptiveness, interactiveness and regularity. Adaptiveness can be achieved when the training program and training materials reflect the properties the test and the characteristics of the audience. Several studies on ineffective training programs have been reported (Crow, 1957; Warmke & Billings, 1979). These studies suggested that the practicality and appropriateness of the training workshop materials and methods should be considered as one of the keys to the success of the training programs. In order for training to succeed, more accurate analysis about rating environment is required, and appropriate guidelines or training methods

should be developed. A training program designer needs to carefully analyze the group of raters at the planning stage in order to achieve adaptiveness and to provide specialized information for the program (Bernardin & Walter, 1977; Borman, 1979; Choi, 2002; Cooper, 1983; Shohamy et all., 1992; Weigle, 1994a; 1994b).

A systematic training program should also be interactive. This means that a training program should offer an opportunity for raters to share their rating problems with others, and share their impression of the rating procedures with test administers by means of an institutional survey or self-reporting. Figure 7 clearly shows the function of the training program, presenting interactiveness in each component. The advantage is that both the training program and the individual raters can get information about the rating process through an iterative feedback process about the usefulness of the training program and individual raters' rating performances based on the interim outcomes of the assessment activity.

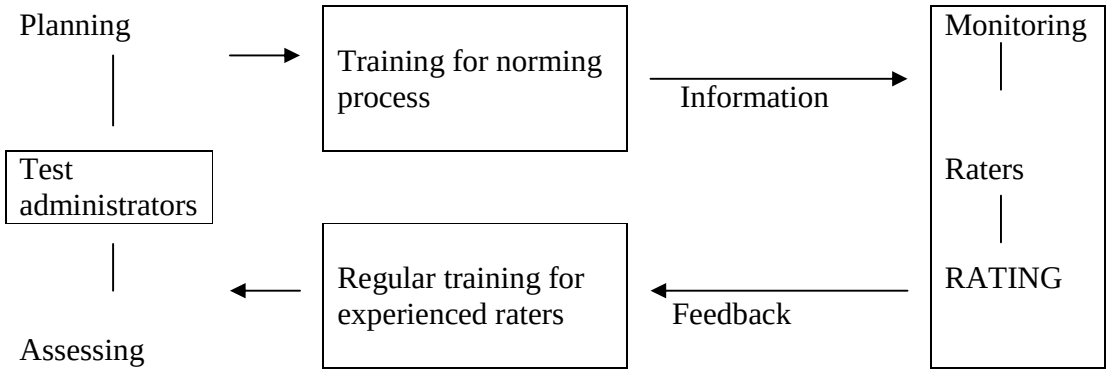


Figure 7. The interactiveness between training program and raters. (Jang, 2006)

Bernardin's and Walter's (1977) results showed that the quality of rating might be influenced by practice, internal assessment and feedback from a professional trainer. It is obvious that more practice and feedback, as well as well-designed scales are key factors

for successful rater training. Charney (1984) suggested discussions with peer raters, and also emphasized the importance of active interaction between a trainer and a trainee, such as practice and feedback. Latham et all. (1975), Levine and Butler (1952), and Wildman, Erickson and Kent (1975) supported these results. A training program including rating practice and feedback from others was more effective than a lecture-only or discussion-only group. It is likely that the interactiveness at the program level and the individual level helps improve training effectiveness and is therefore considered a critical feature in the development of a standardized training program.

Finally, regularity is an essential feature for a standardized training program since rating quality (rater reliability) should be controlled within the training program. In practice, raters are often not provided with a regular training session, or raters participate in brief training programs right before rating. Lumley and McNamara (1995) explored how long raters maintained their self-consistency. Their findings suggested that training did not last long. Only one rater out of three maintained his/her consistency over three rating sessions. They suggest that test administrators provide regular re-training sessions to remind raters of the process before conducting their ratings. Through this study, it is evident that raters' performance may depend upon the length of the rating period. Therefore, the implications of these studies are that more practice and feedback via regular training would help raters maintain self-consistency and solve their rating problems. In addition, the training should provide a separate specialized training program for novices and a continuing program for experienced raters.

For these practical reasons, training materials can be developed to agree on theoretical constructs, and they can be crafted with detailed operationalized descriptors

reflecting the practical situation. Interviews and rating procedures can be accommodated by standardizing the rating method and policy provided by the testing agency. These efforts contribute to improving the effectiveness of a training program, and are good directions toward how standardization should be made.

In addition, after internal research, guidelines or training methods should be elaborated by adding sufficient information on the rating itself and on the rating context. Furthermore, a well-designed training program produces the qualified raters we want, and it may also contribute to obtaining both high reliability and high validity on the test.

True score construction of prototype samples for rating. As well as the development of valid rating descriptors, representative prototype sampling could be considered a feature to be standardized. Standardization supports the true score construction process for scoring reliability. Based on true score measurement theory, the true score can be defined as the expected observable score, although it cannot be directly measured. Practically, the true score can be identified through an observed score, which is directly measurable. Concepts and bias causing measurement errors may change the true score (Goodenough, 1950).

Prototype images in the long-term memory might help maintain reliability, although there is no empirical evidence of this claim. Sulsky & Day (1992) have argued that lack of awareness of unrepresentive sampling for each scale causes more problems in the training context. Since prototype samples help raters understand the rating focus as it relates to rating scale descriptors, they are a strong candidate for enhancing training effectiveness.

Our interest lies in how we can practically obtain “true” scores (as defined in reliability theory), which are not directly observable, from actual testing situations. In addition, along with prototype samples, true score ratings or expert ratings for prototype samples should be carefully chosen in order to evaluate assessors’ performance, because true scores are a ruler for estimating rating accuracy and rating error (Borman, 1975; 1978; Sulsky & Balzer,1988). For example, Furneaux and Rignall (2007) used standardized scores to estimate differences between predetermined scores and individual ratings.

There are several ways proposed for determining true scores for prototypes. In most situations, “true” is defined as “expert” or “official”. True scores for prototype samples can be collected by discussion and agreement across holistic and analytical performance dimensions. True scores can be obtained by scores of experts. Another way to arrive at a true score is to compute the score based on a consensus of a rater group in terms of intra-class correlation (Borman, 1978; Furneaux & Rignall, 2007; Sulsky & Balzer, 1988; Roch & O’Sullivan, 2003). The implication can be found from the study of Haswell (1998) that raters may be to mark the degree to which a prototype matches the rating scale descriptors.

Standardized rater training focus. Borman (1978; 1979) has discussed that training format and method determine interview/rating focus, and might affect raters’ performance. Borman (1977) and Bachman (1988) have also suggested an early concept of standardization, in that raters’ different performance dimensions need to be structured by standardizing the test procedures. In another study (Borman 1979), the focus of training was to standardize the observation of behaviors, and to train raters to understand

a frame of reference and the relative importance of assessment dimension. Borman (1979) suggested a training model to enhance inter-rater agreement as following:

> Now what kinds of training might increase inter-rater agreement in performance ratings, according to this view of the rating process? First, training focused on standardizing the observation of behavior would be important. Second, the model emphasizes the importance of teaching raters a common nomenclature for defining the organizational or societal relevance of the behavior that is observed. Inter-rater agreement should be reached regarding the relative importance of different kinds of behaviors as contributors to effective performance…(p.418).

Based on these studies (Borman, 1977; 1978; 1979; Feldman, 1981; Pulakos, 1984), it would be expected that the standardization of test procedures would be a means to enhancing training effectiveness.

Arvey and Murphy (1998) have discussed that the collective view of rater reliability has expanded from a narrow evaluation of rater performance to consideration of the rating context (the entire rating/scoring system). In order to compensate for the disadvantages of a training program focus on only one aspect many researchers have suggested that a training program should incorporate properties of all four types of training program (Arvey & Murphy, 1998; Roch & O'Sullivan, 2003).

Woehr & Huffcutt (1994) discussed the effectiveness of the different types of training program. Performance dimension training was moderately effective at reducing halo error, but less effective with respect to increasing rating accuracy. Frame of reference training (FOR) training appears to have the most positive effect on increasing rating accuracy. The focus of FOR is to set up common standards for tasks, and the dimensional system is used when making performance judgments. Behavioral observation training with a focus on improving observation skills and correct or incorrect rating distribution showed a positive effect on both rating and observational accuracy.

Rater error training was moderately effective at reducing halo error and somewhat less effective with respect to leniency. Bernardin & Pence (1980) compared error-focused training with accuracy-focused training in their study. They suggested that error-focused training was somewhat effective for leniency and halo errors, but less effective in increasing accuracy.

Research has supported the idea that a new model could be identified by incorporating the benefits of the several types of training (Smith, 1986; Spool, 1978; Sulsky & Balzer,1988). Bernardin & Buckley (1981) suggested that FOR is a good model to acquire valid ratings and reduce rating errors such as halo or leniency or severity by changing a scoring pattern. Since FOR training provides a common standard for assessors, individual eccentric scoring patterns are easily identified in comparison with a standard and expert ratings. For example, in the training session, this can be achieved by making a comparison between individual scoring and expert ratings (true scores), which would be helpful for assessors to be aware of whether their scoring is lenient or not, and how the assessor perceives the standards (rating scale descriptors).

It has been suggested that a new training program could be formulated based on the basic structure of frame-of-reference training by adding the advantages of the other three training models, since many studies support the effectiveness and advantages of the frame-of-reference training method (Athey& McIntyre, 1987; Feldman,1981; McIntyre, Smith & Hassett,1984). In addition, the cognitive rating process model should be considered. Borman (1978) has discussed a rating process model with three steps: observation, evaluation, and judgment. He suggested that when assessors can handle these three stages, rating might be more reliable and valid. Strategies of information

organization or structuring of a global impression or oversimplification, underestimation, and overestimation might make rating more accurate or more reliable. Lee (1985) agreed that categorization is a process of organizing empirical evidence and aids raters in integrating information based on its meaningfulness.

On the basis of these three salient characteristics of this training method, three principles of frame-of-reference training are combined with the advantages of the other three training programs. Feldman (1981) and Pulakos (1984) suggested a generalized information processing model to be standardized: focus of observation, categorization, and judgment based on evidence integration. The principles are (a) accurate observation of information (examinees' performance), (b) accurate selection, (c) accurate categorization, and (d) accurate judgment (evaluation) (Eckes, 2008). It is expected that frame-of-reference training could contribute to the enhancement of the decision-making process using inputs which the assessors receive from the rating context and examinees' behavior.

Accurate observation and selection. In training, observation and information selection processes are standardized by attending to meaningful evidence, rather than simply training raters to recall examinees' behaviors observed. Raters would be trained to accurately observe the communicative competencies of the examinees and to collect separate pieces of evidence from multiple perspectives by observing multidimensional behaviors, rather than reliance on a single aspect of behavior. Accurate evidence gathering could be a stepping stone in the decision making process for the final scoring.

In terms of information observation and selection, Feldman (1981) found that raters tend to be more lenient when paying attention to the positive aspect of behaviors

observed; however, they are likely to be harsher when recalling more negative evidence of behaviors. Bernardin and Buckly (1981) detected that raters tend to more readily remember negative information than positive information. The results suggested implications about what kind of information a rater would pay attention to in the training session. Based on this study, both positive and negative aspects of examinees' behavior are given not only to prevent rating errors such as leniency and harshness, but also to provide a richness of information (Pulakos, 1984). Both positive and negative evidence might be helpful in distinguishing the boundaries of each rating scale level (e.g. between low level, and middle level, or between middle level and high level).

Another implication is that analytic scoring methods for accurate observation and information selection can be suggested for rater training programs, particularly for a novice rater group. Analytic scoring, which looks at more specific and detailed information, seems to be helpful for less experienced raters, because raters need to retain more pieces of evidence separately and to more accurately observe the multidimensional examinees' behavior. McIntyre et all. (1984) provided a counterexample with the view that analytic scoring is not statistically meaningful for the enhancement of rater performance.

Raters are able to distinguish what is an appropriate and meaningful source of information from the multidimensional behaviors. Training about accurate observation of examinees' behavior using positive and negative information and analytic scoring methods is one way not only to avoid information loss but also to enhance raters' ability to accurately select information (Cardy & Kehoe, 1984; McIntyre et all., 1984).

Accurate categorization. Correct categorization is also an important feature affecting the final decision process. Categorization, one of the benefits of frame-of-reference training, could contribute to the accessibility of information by reorganizing the multiple sources of information received. Pulakos (1984) suggested that the central component of the training model is categorization, as it serves to link together the other processes by providing cues for the rater to be able to pay attention to:

> Specifically, the categories in use by individuals guide their attention to particular stimuli while largely ignoring others, and form the basis for subsequent categorizations and recall. ……Hence, to the degree that rater training was focused on creating or imposing a category system to facilitate attention to, storage, and recall of relevant ratee behaviors, performance appraisals should be more accurate, similar idea have been expressed by something. (p.582)

Raters in the assessment context are asked to correctly classify the examinees' behavior into proficiency levels. For instance, raters are provided with two different types of information: rating scale descriptors, which is standardized information from a particular testing agency; and examinees' behavior to be scored. Raters do not simply recall everything they have observed from the two types of input when evaluating the appraisal performances (Feldman, 1981; McIntyre et all., 1984; Sulsky & Day, 1992). "Representativeness is the degree of correspondence between samples scored and a rating category" (Tversky & Kahneman, 2002; Kahneman & Frederick, 2002). After collecting information, categorization plays a role to link what they have observed and what they have memorized for the judgment (evaluation) process (Woehr & Feldman, 1993). Categorization can be taught by establishing prototype images for each proficiency level based on particular rating scales (Arvey & Murphy, 1998; Bernardin & Buckley, 1981; Cooper, 1983; Feldman, 1981; Roch & O'Sullivan, 2003; Sulsky & Day, 1992). Sometimes, categorization helps raters recall detailed information about a global

impression, and a prototype image of each category. Categorization can be achieved by providing appropriate prototype samples across all the levels of examinees' performances during training (Sulsky & Day, 1992). Hauenstein and Foti (1989) discussed the importance of the role of prototype samples in categorization.

Some people might prefer to make a global impression based judgment, while others might prefer an analytic based judgment. However, Lord (1985) has argued that more accurate information might be retrieved based on correct categories pre-existing in their perception. Raters sometimes misinterpret their task and fail to select the decision process due to intuition based reasoning, rather than rule governed reasoning (Tversky & Kahneman, 2002; Kahneman & Frederick, 2002). Conversely, Feldman (1981) has argued that categorization might produce systematic errors (leniency, harshness, score range restriction, and incorrect scoring compared with expert scoring), rather than random errors which cannot be quantified. In spite of the advantages of categorization, this possibility cannot be ignored.

Accurate judgment (evaluation). One of advantage of frame-of-reference training occurs at the judgment stage, as compared with other training methods. Thornton and Zorich, (1980) suggested that judgment processes include categorization, integration, and evaluation of information. Both properties--observation and evaluation--will be simultaneously considered for designing the training program.

Roch and O'Sullivan (2003) asserted that it is likely that one source of rating problems is incorrect categorization, rather than recalling incorrect information. Categorization seems to encourage the simplification of information by forming a global impression image of a particular examinee's performance (Lord, 1985). This seems to

moderate the possibility of incorrect information recall (Slusher & Anderson, 1987) although there is no empirical evidence on a causal relationship between memory and judgment (Woehr & Feldman, 1993).

Cardy and Kehoe (1984), and Woehr and Feldman (1993) agreed that memory effects seem to be negligible in the judgment process once raters are familiar with the rating process, since memory tends to decrease over time. In practical situations, raters could neither remember all the examinees' behaviors nor recall facts they had observed. It is likely that collecting information would be an analytical exercise, while judgment seems to be made by taking more holistic approach. This means that novice raters tend to rely more on their short term memory, but experienced raters might have a different cognitive process—a holistic approach based on overall impression.

Standardization of training content materials. Training program/materials should be crafted with sets of standards from broad vision statements to precisely indicate what raters are expected to know, what they should be able to do, and in what contexts they are expected to demonstrate their proficiency. In the training program, training guidelines and relevant materials related to the test itself, as well as appraisal procedures need to be standardized. Training content can be standardized through analysis of the rating system, consensus on creating guidelines, and empirical data collection. Training guidelines deliver a lot of information; empirical data analysis and consensus process are particularly necessary in the selection of meaningful information, the development of a rating scale (including a global scale and assessment criteria) and the construction of prototype samples.

Standardized training methods. The training materials are designed to reflect clearly how easily raters reach a consensus and how they sustain self-consistency over time. In addition to this, a qualitative approach for in-depth understanding, and observations--reviewing and comparing their ratings—is necessary for enhancing rater quality.

Rater reliability can be determined by talking about rating results during and after training—how closely rater expectations and assessments match—by interacting with individual raters. Mutual feedback from the trainer and the raters is needed. Bernardin and Walter (1977) asserted that if we want to benefit from training, then discussion of the concept of rating errors, analytical observation, and practice with rating scales is necessary. Spool (1978) reviewed studies related to issues of rater training from a historical perspective and introduced specific training methods, length of workshop, and workshop activities, all of which might affect training effectiveness. For instance, lecture was less effective for leniency, accuracy, and halo errors. However, practice and feedback were effective not only at reducing halo error but also at increasing rating accuracy. Bernardin and Walter (1977) found that group discussion was effective for reducing halo and leniency. It is suggested, therefore, that a combination of lecture with practice and discussion sessions is a possible way to facilitate learning during the workshop.

Standardization of training procedures. As well as a sufficient lecture session, norming sessions for rating practice are important to the quality of rating. Norming sessions provide assessors with an opportunity to acquire information about how to rate, based on the following questions: (a) how can I observe the target audience? (b) how can

I select relevant evidence without information loss? (c) how can I make correct categorizations? and (d) how can I make a correct judgment? Parkes (2007) supported this view that prototype sampling can be used in paired group ratings, and during the practice session, group consensus activities should be conducted.

Group activity. Group rating during practice based on consensus might enhance the raters' understanding of the entire scoring system by mitigating rating logic. Johnson, Penny, Gordon, Shumate and Fisher (2005) proposed a "discussion model" to increase the exact agreement percent among raters. The rating results from three different groups--individuals, the discussion group, and the expert rater group (as anchored scores)--were compared in terms of both analytic and holistic scoring methods, and data were analyzed using descriptive statistics, t-tests, and the Spearman correlation. It was found that rating scores by discussion showed a relatively higher correlation with expert scores than those of individuals, which suggests that the use of the discussion model might be a significant method to improve the percent of exact agreement. This kind of comparison yields the insightful implication that, when group activity is adopted in training situations, changes in rating occur, raters better comprehend their rating tasks, and raters have the opportunity to increase their awareness of how they rate.

Elder, Knoch, Barkhuizen and Randow (2008) have investigated the effectiveness of feedback on the performances of individual raters during continued mini-training programs after online workshops. Difference in severity was eliminated, and consistency was improved. Moreover, raters showed a positive attitude toward its usefulness during the feedback session. This group consensus based activity might be effective not only for elaborating the raters' judgment processes, but also for eliminating distinctive internal

standards of individuals. In addition, this type of practice helps raters make their interpretation of the rating scale more explicit. Most scale descriptors and test guidelines encourage raters to take a look at the positive aspects of the examinees' performance by providing "can do" statement descriptors. Conversely, during training, more input ("cannot do" statements) beyond "can do" statement descriptors need to be provided for better understanding of the rating scale. Raters should be trained to distinguish among the major proficiency levels and sublevels.

An implication for training was found from Eckes' study (2008). He suggested two distinctive training models: "behavior-driven (or bottom-up) training" and "schema-driven (or top-down) training" (p. 179). The behavior-driven model clearly distinguishes three cognitive phases and provides a training focus for each step. Meanwhile, the schema-driven model takes a more holistic approach rather than following sequential steps. This approach seems more ideal for adopting a holistic scoring method, and more experienced raters tend to use the top-down model. These findings are relevant to the rater training program in that the bottom-up style model is likely to fit better for inexperienced raters and retraining raters. However, if a holistic scoring model is our destination, a combination of the two models is necessary in the training workshop.

Furthermore, the time duration for the workshop and training procedures is also critical to the learning process. Time duration can play a role in determining training effectiveness (Spool, 1978). Training sessions less than two hours long did not ensure training effectives. The most successful training lasted between 3 and 14 hours. A workshop session one or two days long would be sufficient to convey all of the content

for learning and practice. The training workshop for the ACTFL OPI, for example, runs from three to five days for the full rating scale workshop (http://www.actfl.org).

Feedback for raters. A study by Furneaux and Rignall (2007) investigated changes in rater perception using written reports, and found that the decision-making process seemed to be dynamic. Shaw (2002) also supported this point, showing that raters elaborated their rating behaviors over time. This implies that training can enhance rating performance in a short time, but sometimes it can have a negative impact on raters, depending on what type of feedback is given to the raters. Wiggleworth (1993) provided individualized feedback for speaking raters based on consistency and bias analysis. Lunz, Wright, and Linacre (1990) suggested that individualized feedback seems to work better (Cited in Shaw and Weir, 2007). O'Sullivan and Rignall (2007) have studied how feedback affects rater performance by investigating their consistency and perception via survey. They provided individual verbal descriptions of the results of bias analysis, a z-score of consistency, and severity levels using graphical information. Based on raters' responses, it turns out that raters prefer to have graphical descriptions as well as verbal reports. Rater groups that received feedback showed improvement in their scoring reliability and perception. Nonetheless, Elder et all. (2008), O'Sullivan and Rignall (2007), and Wiggleworth (1993) suggested that onetime feedback does not last long, and a more regular feedback system is necessary for raters to maintain a high quality rating performance.

Measures of Training Effectiveness

Standardization of measures of the entire training workshop at the institutional level. Program theory evaluation provides the overall direction of this study--not only how the program is understood, but how the program actually works. In particular, Weiss's (1998) program evaluation theory (Personal communication with Prof. Jennifer Greene, 2007) and the context adaptive model (CAM) for evaluating educational language programs proposed by Lynch (1996) are applicable to this study. Evaluation can be defined as a structured and intended assessment tool with explicit evaluation standards to compare outcomes of the program and judge whether or not the various kinds of educational programs are functioning well for the recipients. Lynch (1996) and Weiss (1998) have defined the phrase "program evaluation" below.

> Evaluation is defined here as the systematic attempt to gather information in order to make judgments or decisions. Program, in general, it tends to evoke the image of a series of courses linked with some common goals or end products. A language education generally consists of a slate of courses designed to prepare students for some language related endeavor." "I will use program to refer to any instructional sequence, such as multilevel English as a second language curriculum, and a foreign language teacher-training workshop is self-assessed by students in a language lab (Lynch 1996, p.2).

> Evaluation is the systematic assessment of the operation and/or the outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy (Weiss, 1998; p.36-39).

The primary focus of evaluation is at the program or institutional level, in spite of the fact that the boundary between assessment at the individual level and at the program level seems to be ambiguous. It is a fact that the activities of program's participants are strongly related to the interim or end results of that program. So the results of the

assessment of individual performance could frequently be used to decide the degree of improvement in the quality of the program (Lynch 2003; Weiss, 1998).

Weiss (1998) introduced the basic concepts of program theory and implementation theory. These theories describe the overall evaluation process step by step: (a) determine the overall goals of the evaluation; (b) adjust or specify research questions after examining the evaluation context with key stakeholders from the testing agency; (c) set up criteria for evaluation and to build program input and activities; and (d) anticipate the interim program outcomes and ultimate desired product.

On the basis of the definitions for program evaluation of Lynch (1996) and Weiss (1998), the benefits of adaptations of program evaluation can be identified. First, program evaluation seems to take a more systematic approach to using the appropriate philosophical paradigm. The philosophical perspective guides the design of evaluation research using quantitative, qualitative or mixed methods. Second, program evaluation considers not only the operational process of a particular program, but interim short term outcomes and long terms products in the end. An understanding of the current environment of a particular program assists in correctly interpreting the short-term and long-term outcomes. Another point is that more explicit criteria for comparison and judgment are necessary based on congruence among stakeholders. Program goals and different expectations of stakeholders are synthesized while carrying out the evaluation research. In addition, it is helpful to enhance the objectivity of the judgments. Finally, in terms of the utility of evaluation outcomes, the results of the evaluation contribute to enhancing the quality of a program or individual's performance by periodic reporting and

publishing of the reports. So evaluation research seems to have more powerful washback effect on the educational system in comparison with general research.

Lynch' context adaptive model for evaluation. Lynch (1996; 2003) has discussed the context adaptive model (CAM) for program evaluation as it is applied to language education and testing. This model proposes that the entire evaluation procedure should reflect the specific concerns of stakeholders and should be designed depending on the context of the particular program being evaluated. In line with this perspective, program evaluation procedures could be considered a possible evaluation model for realizing the standardization of a rater training program by covering the entire scope of evaluation. Lynch's model is context adaptive--responsive to the evaluation environment—because it provides seven phases for evaluation research from analysis of the audience and elaboration of the evaluation goals as a result of two ways of communication with stakeholders and the evaluation context (see Appendix C and D).

Lynch's model was modified for the EPT rating environment. The revised model maintains the seven major phases, but the research activities at steps 4, 5, and 7 were revised to accommodate conducting a program evaluation of the EPT rater workshop, since that requires more interaction with stakeholders. Interim outcomes from each phase are necessary. In Lynch's model, step 4 was originally designed for data collection, but in this study, it was eliminated, after careful consideration. Step 4 became "Standardization of the EPT training workshop." Step 5 became "Implementation of training workshop," and step 6 was changed as post-rating and data analysis, but is the almost same as Lynch's original model. Figure 8 exhibits the distinguishable features of Lynch's model and the revised model used in this study.

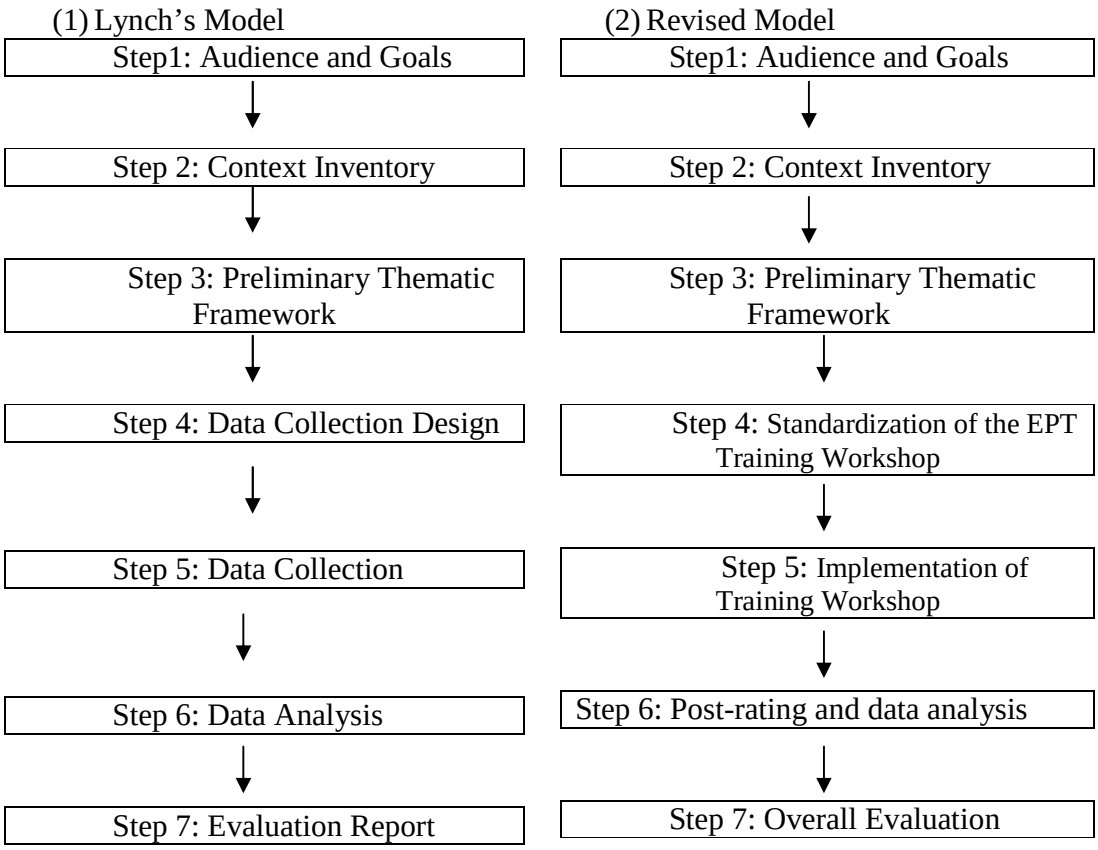


Figure 8. Lynch's context adaptive model (CAM) and revised model.

General measures for rating variability. In this stage, there are two critical issues to be discussed. One is the determination of how evaluation research is characterized--for instance, formatively or summative. Second, the primary attitude toward how to design the research should be decided. Which approach is more suitable for evaluation research--positivistic approach, interpretive approach, or mixed approach? This is an important decision in the sense that these two efforts help determine specific methods for data collection and analysis.

Evaluation can be either formative or summative, depending on the goals of the evaluation. Formative evaluation is seen as an assessment of the ongoing progress and demands of the participants in a particular program while that program is developing. It

has instructional purposes focus on what individuals have achieved and how well the program is functioning. The final product of formative evaluation suggests some changes for improving the quality of individual performance and the program. The concern of summative evaluation, on the other hand, is high-stakes decisions about the worth of a program--whether the program can achieve its goals and expectations, and whether it is worth financial support. For this study, formative evaluation will be conducted to estimate the effectiveness of the program after a training workshop and curriculum development for raters.

Lynch (1996; 2003) has discussed a positivistic perspective as a paradigm of program evaluation in detail, since this approach plays an important role in determining research design. First, the positivistic perspective is concerned with two things: certainty and generalization. Positivists are interested in accurate measurement and inferences from the program, and they try to identify the relationship between causes and effects. They are also interested in generalizing the evaluation results to other evaluation settings. This approach suggests quantitative data gathering during short or long term periods. A certain format of treatment is provided for a program group to estimate the effectiveness of treatment. This approach tends to prefer the use of experimental or quasi-experimental design, and the outcomes of the program group and comparison group are compared to see the impacts of treatment. Research design can be determined depending on the evaluation context.

Another advantage of the proposed training program is its ability to monitor raters' professional development by evaluating their performance. Bernardin and Buckley (1981) suggested that the frame-of-reference training system might be effective for

training and changing raters' idiosyncratic scoring patterns by monitoring their scoring tendencies. It is suggested, therefore, that this monitoring stage also needs to be standardized to maintain the quality of rating. Several standards for measures of rating skills will be identified. For instance, in terms of psychometrics, due to a lack of research, the biases in rating are not be clearly defined, and the sources of rating bias are unclear, although quantitative methods such as FACETS and GENOVA analysis are used for quality control (Lunz et all., 1990; McNamara, 1996; Shin, 2001; Weigle, 1994a; 1998). Valid rating patterns regarding accuracy (inter-rater reliability), halo errors, range restriction, contrast effects, and similarity effects are also considered for justifying the quality of rating (Hauenstein & Foti, 1989; Roch & O'Sullivan, 2003). In addition to this, measures at the program level should also be considered on the ground of empirical findings.

Accuracy measure. Accuracy can be computed in two different ways: distance measures from the true scores of prototypes by expert raters, and correlation measures, which is the same concept as inter-rater reliability. Roch and O'Sullivan (2003) refer to distance measure as "the average absolute deviation of rating from the true score." It is estimated based on the direct comparison between the two score distributions or ranking order of individual ratings and expert ratings (Borman 1977; Johnson et al., 2005; Pulakos, 1984; Sulsky & Balzer, 1988). Correct rank ordering (a correlation of 1.00 being perfectly sensitive to the expert rating) and the deviation of the mean score between the individual ratings and the expert score is a substantial ingredient in the calculation of the rating error: 0 means perfect, a negative sign means that a rater underestimates

performance, and a positive sign means that a rater overestimates the performance (Hauenstein & Foti, 1989).

Bernardin and Walter (1977) defined inter-rater reliability as the degree of agreement between raters on each criterion. The mean of Fisher's r-to z transformation correlation identifies the similarity of overall rating patters (Athey& McIntre, 1987; Borman 1977; Borman, 1979; Pulakos, 1984; Woehr & Huffcutt, 1994). A high correlation indicates a greater agreement and a smaller SD among the ratings (Saal, Downey & Lahey, 1980).

Score range restriction. Score range restriction is defined as a rater's limited rating pattern or tendencies exhibited in the obtained scoring, which may be lenient or severe in comparisons with the true scores of prototype samples (Athey& McIntre, 1987). Leniency refers to a higher mean rating than the expert ratings, while severity refers to a lower mean rating than the expert ratings.

To say that a rating pattern has a central tendency means that the score dispersion is around the middle scores on the rating scale (Saal et all., 1980). It is obvious that raters prefer to use middle scores rather than using the extreme scores on the rating scale continuum. It seems that lack of confidence in rating might be the main cause of using the middle scores.

Score range restrictions can be computed as the mean difference from each assessment criterion and holistic scores (Athey& McIntre, 1987; Bernardin & Pence, 1980; Murphy & Balzer, 1989; Pulakos, 1984). Skewness, kurtosis and standard deviation are indices for computing the restrictions for the score range (Borman 1977; Murphy & Balzer, 1989). Score range restrictions can be regarded as a source of bias

which might threaten rater reliability, because this is evidence that raters cannot discriminate either between or within examinees' performances (Saal et all., 1980).

Halo error. Halo error (low discrimination or oversimplification) is defined as an individual difference in standard deviation among assessment criteria for multiple dimensions of a behavior being observed (Athey & McIntre, 1987; Bernardin & Walter, 1977). This measure reflects the extent to which a rater considers each dimension separately in rating (Borman, 1979). Saal et all. (1980) discussed that halo error may appear because of the effect of a global impression on the scoring process of each dimension rather than to cautious discernment of each assessment dimension within an examinee's performance. Higher correlations ($r \geq .8$) or lower SD (variance) indicates a higher halo effect and suggests less discrimination across different aspects of behavior than true expert ratings (Bernardin & Pence, 1980; Borman 1977; Murphy & Balzer, 1989; Pulakos, 1984; Saal et all., 1980[4]; Woehr & Huffcutt,1994). Raters are encouraged to assign their ratings by evaluating the various dimensions within a single examinee.

Contrast and similarity effects (error). Contrast error can occur because of ordering in rating, and it might appear as a result of comparing among performances than comparing the performance to the rating standards (Latham et all., 1975). Contrary to this, similarity error may appear when a rater's attitude and/or background (e.g. experience or familiarity with pronunciation) have something in common with the test taker's performance (Bernardin & Walter, 1977). Both errors can reduce rater reliability, since these errors can produce rater-particular examinee group interaction effects. The rating

[4] Saal, Downey & Lahey(1980) suggested to compute the halo error. 1) Fisher's r- to-z transformation is computed. 2) The true dimension inter correlations were subtracted from the observed dimension intercorrelations. 3) The difference scores were averaged, providing a mean measure of the difference between the true and observed intercorrelations across dimensions. 4) To the degree that this average deviated from zero in a positive direction, more halo was present.

patterns can be easily biased for particular examinee groups or assessment criteria, and intra-rater reliability can be eliminated as well as inter-rater reliability, as a result of in rater-examinee interaction.

Generalizability theory for estimating raters' variance component. Validity as well as reliability in rating can be estimated from the amount of variance using G-analysis (Borman, 1975; Lievens & Sanchez, 2007[5]). In the area of language testing, the usefulness of G-theory has been discussed in several studies (Bachman et all., 1995; Lumley & McNamara, 1995; Lynch & McNamara, 1998). G-theory analysis estimates random variance components for multiple facets of examinees, raters, and assessment criteria. This estimate provides insightful evidence for scoring reliability by estimating the G-coefficient, and for rating validity by estimating the amount of variance of each facet. GENOVA analysis is useful to understand the overall effect of any facet and to provide interaction effects between two facets. It yields greater implications about test design, test revision procedures, and overall evaluation of test validity.

FACETS analysis. FACETS analysis has been used to analyze rater reliability in order to look more closely at the level of raters' severity or leniency, individual consistency, agreement rate among raters and any bias patterns between facets (Linacre,1989). Currently, researchers (Bachman et all., 1995; Brown, 1995; Choi, 2000; Kondo-Brown, 2002; Lumley & McNamara, 1995; Lunz et all., 1990; Lynch & McNamara, 1998; McNamara, 1996; Shin, 2001; Weigle, 1994a; 1994b; 1998) are using this model to examine rater reliability issues. FACETS focuses primarily on estimating the performance (or quality) of individual aspects. The advantage of this analysis is that it

[5]The variance component due to competencies represents a desirable source of variance because it indicates discriminate validity across competencies (competencies variance, discriminate validity). Competency x raters refers to inter-rater reliability which it gauges variation in competency ratings across raters.

can simultaneously analyze various facets which are involved in a study (e.g test-takers' ability, item difficulty, rater characteristics, assessment criteria) affecting scores. Furthermore, it also provides evidence for rating scales. Particularly, these specific findings of FACETS can be useful for rater training because they provide feedback about the performances of individual raters.

Chi-square analysis for agreement. In order to evaluate raters' performances during training, a pattern of agreement among raters should be identified. For this study, percent agreement; correlation analysis; Cohen's Kappa Measure, based on quasi-symmetry models; and log-linear (L×L) association models were employed in addition to descriptive statistics, and the results were compared (Agresti,1988; 1996; Agresti & Winner,1997).

To summarize this chapter, a new standardized training program would carefully consider a balance between rater reliability and rating validity by considering anything that could affect interviews and rating. A new rater training program could be designed with regard to this classification, depending on its goals and the needs of the rating environment. In general, the three principles of frame-of-reference training could be reached by a series of activities during workshops: (a) some standardized information ensuring quality, (b) activities during training session for practice and feedback from a trainer, and (c) consistent rationales for judgments from a trainer (Arvey & Murphy, 1998; McIntyre et all., 1984; Roch & O' Sullivan, 2003). The training materials for guidelines and practices should be standardized to offer sufficient information for trainees to be able to successfully meet their rating responsibilities. This means that standardized materials are efficient at increasing the quality of the training program and

reinforcing the learning process. In addition, quantified measures are good indicators of rating validity and reliability. Consideration of both rating validity and reliability could efficiently reduce rating variability. It is possible, in other words, that through these measures, rating validity and reliability may increase because of a change in the raters' rating patterns during training, rather than attending to rating errors (Bernardin & Buckley, 1981). When the training program is standardized, raters receive similar input, thereby ensuring the quality of the information. Most training materials have some limitation as to their ability to deliver the message, but standardized guidelines help reduce the gap between the ideal goals of the training and the actual rating context.

Chapter Three

Research Design and Methodology

In the chapter, the research activities and methods for data analyses at the macro level (analysis of program) and micro level (analysis of individual measures) are described, on the basis of Lynch's modified program evaluation model. Research questions, research procedures and activities, methods for specific measures, and the focus of data analysis at different levels will be introduced.

Rationale for Using Mixed Method Approach

For this study, a mixed method approach was adopted for data collection and analysis. Mixed methods were introduced by Greene, Caracelli, and Graham (1989), and Caracelli and Greene (1993). Greene et all. (1989) stated that mixed methods involve using more than one method to collect and analyze data, both quantitative and qualitative. Kim (2008) identified the appropriateness of using mixed methods for research about language testing, to better portray its complexities, reflect the perspectives of the diverse stakeholders, and increase research validity.

The research phases of the revised Lynch's context adaptive model (CAM) are also related to the evaluation model. This research took a formative evaluation approach in order to provide interim feedback for improvement, rather than a summative evaluation approach which would decide whether the rater training should be terminated or continued. The focus of this evaluation research was the ongoing process of the rater training program; interim/final output was also incorporated at the end of the research. Employment of mixed methods was appropriate for this study, because data were

collected from various sources during the different research steps, so it was necessary to use mixed methods when analyzing these data. In addition, mixed methods yield a more meaningful interpretation, rather than a biased interpretation which might have arisen from using a single method, due to limited findings.

Diverse perspectives can provide a better measure of training effectiveness, and were achieved by combining the results of both quantitative and qualitative approaches. Training effectiveness was estimated at several layers: Interim outcomes of each stage were analyzed and reported: some changes were noted at the individual level; and some changes were evident at the program level. To see the changes in participants' perceptions, interim outcomes at each research step were examined, such as responses to the open-ended questions on the surveys, group activities during the workshop, and raters' individual reflection logs. Field notes from the revision process of the current training materials and workshop implementation were analyzed using qualitative methods. At the program level, interim or final outcomes from surveys, the workshop activity, and findings of post-rating with mixed methods comprised one standard for estimating training effectiveness (see appendix D). Figure 9 exhibits the research design and specific research activities corresponding to major research steps.

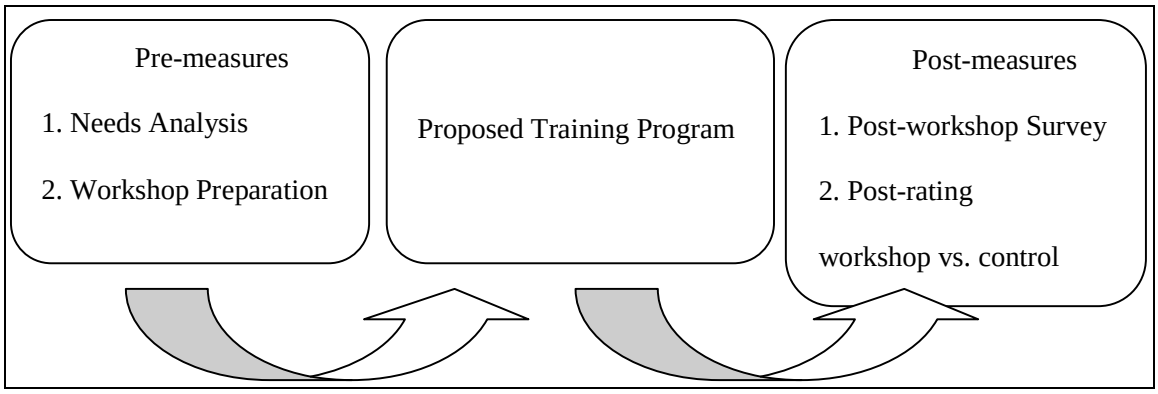


Figure 9. Research design.

The Pilot Study

My early research on rater training served as a pilot study for this dissertation. The pilot study examined changes in the reliability of raters, topics, criteria, and scales in pre- and post-rating sessions, and evaluated how much the training program contributed to improving reliability and validity for the EPT test. Additionally, a post-workshop survey was conducted to see how raters evaluated the training program.

The EPT training program was effective because the raters' performances in the post-rating session improved over their performance in the pre-rating session in both FACETS results and analysis of variance. In terms of inter- and intra-rater reliability, it was shown that the reliability of the raters in the post-rating session improved over that of the pre-rating session. According to FACETS reports, the differences in severity among raters decreased and the percentage of agreement increased. In the pre-rating session, two raters showed misfit and overfit patterns in rating, but the self-consistency of the raters improved since we did not detect any evidence about misfit and overfit rating patterns in the post-rating session. The two topics were almost equivalent in the two rating sessions. However, it was clear that the raters gave a different weighting to each criterion. Content was an important area when raters decided the holistic score, while linguistic expression was least important in both rating sessions. The four levels of the EPT scale functioned better in the post-rating session. Looking at the results of the survey, raters seem to be satisfied with the training program, and they fully understood the EPT rating procedures.

This study focused on the effectiveness of the training program for ESL teacher raters. The effectiveness of training program was reviewed in terms of raters and test format. There were several implications from the pilot study that served as the driving

force behind my dissertation. The first implication was that we need to provide more careful training programs to fit the EPT rating context. Second, the pilot study showed that reliable rating or grading is an important part of teacher training, and that key issues are how to give positive feedback, how to facilitate communication between raters and trainers/supervisors, and how to train raters more systematically. The pilot study also showed that a proposed training program should be designed to solve not only theoretical but also practical concerns. Finally, my experience with the pilot study informed the research methodology of this dissertation, indicating that a qualitative approach would be necessary to fully understand the complexity of rater concerns. While conducting this research, the suggestions of the pilot study contributed to the formation of the main study.

Research Questions

The primary objective of this study was to develop a systematic rater training program for the ESL placement test with a fuller perspective of rater reliability. The proposed training model was based on a review of the literature, especially the revised CAM model by Lynch. Research questions were formulated for each stage of the proposed new model. In order to formulate the research questions, considerations about the test instrument itself, the test procedures, participants' educational training, and contextual effects, such as the characteristics of the stakeholders, the structure of the test battery, and washback effects of the test instrument, were fully taken into account. This study was guided by the following questions:

> 1. What issues (concerns) about the evaluation context do stakeholders perceive? How do raters evaluate the current EPT training program? How do raters perceive the rating methods, EPT scales, criteria, and topics? What are issues in terms of scoring reliability?

2. Which part(s) of the training program can be standardized? How can the EPT training program be standardized?

3. Was the proposed new training workshop successfully conducted? What sources might strongly influence raters' score reliability?

4. How effective is the proposed rater training program, and do all stakeholders agree that the standardized training session is helpful for enhancing their rating performance? Are there differences in the quality of rating before/after standardized training in terms of both classical (inter- and intra-rater) reliability as well as a broadened view of it? Are there differences in raters' performances (rating skills) before/after training with regard to the individual backgrounds of raters (i.e. rating experience and teaching experience)?

Data Collection Procedures of This Research

Step 1: audience and goals.

Setting up evaluation goals. The primary objectives of this study were to evaluate and develop a systematic rater training program for the EPT with a fuller perspective of rater reliability. A secondary goal was to identify the effectiveness of the proposed training workshop based on the stakeholders' perceptions and measures of their performance.

The different participants at different phases from the same population. Identification of the audience and goals of the program is the first step of CAM. It is important to clarify who the audience(s) and stakeholder(s) are since the goals of the evaluation are determined by the interests of the stakeholders (Lynch, 1996; 2003).The target audience of this study was comprised of the people involved in the EPT essay test. The recruitment procedures were carried out in two phases, and the raters and coordinator of the ESL writing courses (trainer) participated in this study.

Different participants were involved in the different phases of this research. They were all recruited from the same population, because they were all ESL teacher raters. Some of them were still serving as ESL teachers during the semester that this research was completed, and were do not. Three different recruiting processes were conducted to meet the research requirements, and participation in each research phase was voluntary. The coordinator of the ESL writing courses (EPT trainer) was newly hired last fall semester, and she was actively involved in this study. She had a lot of ESL teaching and rating experience.

Pre-workshop survey. For the needs analysis, eight ESL teacher-raters who were teaching or had taught ESL writing classes participated in this survey, and the survey was collected in person or via e-mail at the participant's convenience. Based on the analysis of the respondents' personal profiles, the majority of the teachers were female (n=6), while two male teachers participated in the survey. Seven participants identified themselves as non-native speakers of English, and one identified as bilingual.

The participants reported various levels of teaching experience, from one year to more than three years. One participant had had one year of teaching experience; three participants had two years of teaching experience; and four participants had three years or more of teaching experience--two had three years, and two had more than three years. The participants reported similarly varying degrees of rating experience. One respondent was a new rater, having never rated or trained before; one respondent had one year of rating experience; two respondents had two years of rating experience; and three respondents had three or more years of rating experience--two had three years, and one had more than three years. Seven of the eight respondents had participated in EPT rater

training once (n=3) or twice (n=4). Two respondents had had rating experience at another institution, while six had not had any rating experience outside of the EPT.

Workshop program evaluation. Fifteen raters participated in the proposed training program and workshop evaluation. They were all ESL teachers who were teaching ESL writing courses during the spring semester of 2010.

Post-rating session. Six raters were recruited for the post-rating study, and they were divided into two groups. Three participants had already participated in the proposed training program (experimental group), while the other three had not been involved in the training program (control group). Half of these six raters had had more than one year of rating experience, and half of them were new raters with no EPT rating experience. This would be helpful for observing the impact of the rating/teaching experience.

Understanding the evaluation context: EPT essay tests. The data for this study were collected from the ESL Placement Test (EPT) at the University of Illinois at Urbana-Champaign (UIUC). The researcher observed the entire essay testing situation in June 2009, with the permission of the EPT director. The description of the test procedures are as follows. The writing test is a paper-based essay test which takes almost three hours to complete. Test takers have consecutive activities with an ESL teacher and peers, who are sitting next to one another. At the beginning of the test, the EPT procedures and topic are introduced, and test takers are provided with a topic-relevant article. After reading the two-page article, the test takers participate in a 30 minute mini-lecture and discussion with an ESL teacher and peers. There are three topics for the EPT test: globalization, cloning, and animal testing. One of the three topics is randomly chosen, and an ESL teacher gives a short lecture session defining the basic terms using an overhead projector,

and giving directions about how to collaborate with peers. The teacher also briefly explains the scoring guidelines to the test takers. The test takers make a first draft of their essays, and then they review it with their peers. Finally, the test takers individually write a final draft of their essays. In this stage, the test takers have 60 minutes to write a two- to three-page essay.

The operational EPT rating environment. After collecting the essays, the EPT administrator distributes the essay the same afternoon, and normally two raters participate in the marking process. Rating essays is one of the responsibilities of the ESL teachers, assigned on their duty rosters. Two raters are asked to give a single holistic score for each essay. The holistic scores consist of four levels which are related to the ESL writing courses: too low; ESL 500 (for graduate students) or ESL 113 (for undergraduate students); ESL 501 or 114; and exempt (or ESL 115). Next, if there is a discrepancy between the scores assigned by the two raters, discussion is required to reach an agreement. If the two raters do not reach a consensus through discussion, then a third rater decides the score. Although analytic scores are not used in operationalized rating, the EPT trainer offers five assessment criteria in the workshop: focus, support/elaboration, organization, convention, and integration.

Step 2: Context inventory (needs analysis). A deeper understanding of the context of a particular program is important to program evaluation. In this second step, the characteristics of the program dimensions to be evaluated and its setting are clarified and identified to the fullest extent possible, relying on the collection of detailed information. To understand the evaluation environment and resources, a needs analysis was conducted as a pre-evaluation. Moreover, some constraints of evaluation and limits

emerging during the following evaluation stages were also identified. Needs analysis based on document analysis and surveys provided knowledge of how well the program worked; what issues need to be identified; and what the beliefs or expectations of the people involved in the program were, regarding the relationship between rater reliability and the effectiveness of the training program (see Appendix E, F G and K).

Step 3: Preliminary thematic framework. On the basis of the analysis of the audience, the evaluation goals, and the evaluation context, some evaluation arguments and critical issues arising in the current program can be provided in the preliminary theoretical framework. The framework makes the focus of the evaluation clear and assists the evaluator in carrying out research design, data collection and analysis. Detailed content materials, activities, training methods and institutional expectations in terms of reliability (agreement/consistency/quality control) were determined based on a review of the literature surrounding training theory and language testing theory.

In Step 3, based on the interim outcomes of Step 2, the researcher confirmed the findings of previous studies so that an appropriate workshop framework could be formulated (see Chapter 2). For this study, three major theories--evaluation theory; theoretical concepts from the training theory in human resources; and theory of language testing, including issues of measurement, contributed to the design of the workshop framework.

Step 4: Preparation of the EPT training workshop. In consideration of the theoretical paradigm and the practical demands, a discussion with the EPT trainer about how to achieve standardization of the rater training program, including training input and

activities and institutional support, was necessary. The literature review and the data collected at Steps 1 and 2 were use to revise the existing workshop materials.

To prepare the workshop, the researcher met with the EPT trainer and collaborated on developing the training workshop, revising the EPT training curriculum and workshop plans as needed. The researcher shared her knowledge about the literature surrounding rater training with the trainer by providing a short summary of the findings of previous studies and the results of pre-rating. In collaboration with the trainer, the researcher decided the workshop schedule and how to standardize the workshop materials and procedures (see Appendix I, J and M).

Step 5: Implementation of the training workshop. This section provided evidence about how well the workshop plans had been implemented. Through rigorous discussions between the researcher and the trainer based on interim findings from Steps 1, 2, and 3, a clear picture of the workshop was presented, and the training workshop was conducted by the EPT trainer. Both experienced and new raters participated in the training. The training workshop was expected to last approximately three to four hours. The training program provided a short familiarization and norming session to understand the rating context. In addition, raters were asked to complete a survey to evaluate the revised training program (see Appendix H, L, N, O and P).

Step 6: Post-rating session: experimental design. The post-rating session was designed to investigate training effectiveness by estimating the raters' scoring performance. The post-rating session was conducted one week after the workshop, from January 18th to 30th, 2010. A simple experimental design (workshop group and control group) was proposed for this research in order to evaluate the effectiveness of the training

program (Wexley & Latham, 2002). In this proposed design, one rater group was the workshop group with workshop inputs, and participated in the onsite rater training session. The control group did not participate in the workshop (no input). One week after the workshop, the same essay package was given to the two groups, and the same rating procedures were followed. The post rating session lasted two weeks.

Six raters voluntarily participated in the post rating, so three raters were assigned to each group. The scoring reliability of the six raters was estimated using a quantitative approach and the results of the measures were compared to see how effective the training workshop had been. Raters were asked to score selected essays in terms of both holistic and analytic scores, and to record a written reflection log in order to note reasons for difficulty in determining scores, rationales for the revision of scale descriptors, and their scoring strategies. Rating guidelines, such as a rating sheet and reflection log were designed and provided by the researcher. Scoring results were collected by e-mail. The rating results from the post-rating were analyzed using quantitative methods. The analysis focused on rating accuracy and reliability. Raters submitted a retrospective written report to shed light on their decision making processes (see Appendix J, Q and R).

Step 7: Data analysis for the overall evaluation. At this stage, based on an understanding of the needs of the evaluation, the different kinds of data collected from the interim outcomes at each research phase were incorporated. After ensuring the appropriateness of the data collection procedures, the data were analyzed and interpreted. The findings were used to enhance the quality of rater performance and of the proposed systematic rater program. In the final stage, meaningful findings were documented in the form of a short evaluation report and communicated with an audience.

Data Analysis

In order to answer the four research questions, data were collected in the steps described above.

Data collection and analysis for question 1. A needs analysis was necessary to understand the program context. There were three different sources of data: the current training materials, the test specifications, and a survey of the stakeholders.

Documents analysis. Documents were collected from the website (https://netfiles.uiuc.edu/fgd/www/ept_bulletin.pdf), and some materials for internal use provided by the coordinator of the ESL writing courses and EPT administrator were analyzed. The EPT bulletin provided information about how to register and prepare for the test. It also included test procedures, descriptions of test tasks, and ESL course guidelines. Scoring guidelines, provided by the EPT administrator, were critically reviewed in order to revise the training program. In addition, the current EPT training program was analyzed based on open-ended questions in the survey, documents analysis of training materials, and the researcher's past observation of the training workshop.

Pre-workshop survey. The surveys were administered to the raters and the trainer. The survey consisted of two parts: questions using four-point Likert scales, and open-ended questions. For raters, the questions covered various topics related to the EPT test, scoring procedures, the current training program, and trainer evaluation. For the trainer, some open-ended questions about designing a new training program were added.

Both qualitative and quantitative approaches for survey analysis were used. Because of the small number of participants (n =8), descriptive statistics were used in

analyzing survey questions, and the responses to the open ended questions were categorized based on themes/topics and analyzed.

Data collection and analysis for question 2. Data were collected from two sources: my notes from meeting with the trainer and the training materials.

Meeting notes. The researcher had biweekly meetings with the trainer from October, 2009 to December, 2009, and three more meetings in January, 2010, before the workshop. The researcher had made field notes to summarize what aspects of the workshop curriculum needed to be changed, with rationales. After discussion with the trainer, some changes were recorded in the form of field notes.

Content revision. On the basis of the results of the discussion, the researcher provided a summary of the rationale for replacing training content, training methods, and prototype essays. This included a summary of the findings of the pre-survey and some suggestions for revisions based on the literature.

Data collection and analysis for question 3. There were two data sources: observation of the workshop and the materials and products of the workshop.

Observation. The researcher observed and took notes on the entire two workshop sessions, and they were audio-recorded. After the workshops, the audio-recordings of the raters' discussions were transcribed and analyzed using a qualitative approach in which their dialogue patterns were categorized according to themes.

Rating practices. All materials and products of the activities were collected and analyzed using simple descriptive statistics.

Data collection and analysis for question 4. Data were collected from two sources: a post-workshop evaluation survey, and post-rating results.

Post-workshop evaluation. A survey was provided for raters at the end of the workshop. The survey asked raters to mark how much the workshop had helped refresh their memories and resolve their problems. Descriptive statistics were employed for this analysis, and responses to open-ended questions were categorized and summarized.

Post-rating session. A post-rating session for the writing assessment was conducted. Six raters were involved in the post-rating session. Two rater groups were asked to follow the same rating procedures, and to write a reflection log to explain the rationale for their scoring and decision-making processes. Quantitative methodology was used in analyzing the rating data, including descriptive statistics to reveal rating patterns, FACETS and G-theory to look at rater reliability, and Cohen's Kappa analysis to examine agreement. In order to analyze the reflection logs, responses were summarized and categorized. Figure 10 shows an overview of the research design and methods.

Chapter Summary

This chapter describes the overall research structure, procedures, and methods for data collection and analysis in detail. Quantitative and qualitative data were collected over the course of one semester. The effectiveness of a proposed rater training program was estimated from a wider perspective based on four research questions. The findings will be described in the chapter 4.

Step(s) of revised CAM model	Research Question	Purpose	Data Source(s)	Participants	Analysis
1 & 2	1. What issues (concerns) about the evaluation context do stakeholders perceive?	Understanding the evaluation context and identifying the issues or concerns of stakeholders	• Pre-workshop survey • Internal EPT Documents	Eight raters with EPT rater training experience and the EPT trainer	Descriptive statistics and document analysis
3 & 4	2. How can the EPT training program be standardized?	Identifying the appropriateness of the proposed new training workshop	• Meeting notes	Researcher collaborated with trainer to elaborate training materials.	Content analysis to modify the training materials.
5	3. Was the proposed new training workshop successfully conducted?	Identifying successful implementation of the training workshop	• Observation of the workshop • Analysis of workshop activities	Fifteen raters who were currently working as ESL teachers	Observation notes, raters' marking and group discussion, using mixed methods. Descriptive statistics were adopted for the surveys.
6 & 7	4. How effective is the standardized rater training program?	Evaluating the effectiveness of the proposed EPT training program	• Post-workshop survey • Investigating rating split rate • Post-rating data including refection log	Six raters	Descriptive statistics, FACETS and GENOVA, and Cohen's Kappa were used for estimating rater reliability.

Figure 10. Overview of the research design.

Chapter Four

Results

This chapter presents the results of the four research questions which were formulated based on the phases of Lynch's program evaluation model. This study adopted quantitative and qualitative approaches to gather evidence from a variety of sources.

Findings of Research Question 1

Research Question 1:

1-1. What issues (or concerns) about the evaluation context do stakeholders perceive? How do raters evaluate the current EPT training program in terms of EPT scales, criteria, the rating methods and topics?

1-2. What are issues/concerns of the EPT rating environment?

A pre-workshop survey and document analysis were employed to answer the first question. The pre-workshop survey was conducted in the fall of 2009 to gather evidence about the evaluation context, and to identify stakeholders' concerns about the EPT rating environment. Eight ESL teacher-raters voluntarily participated in the survey session, having been recruited from the ESL writing courses. The survey consisted of two parts: closed-questions using four-point Likert scales, and open-ended questions covering various topics related to the evaluation of the EPT test, the scoring procedures, the current training program, and the trainer. The respondents were asked to mark the extent to which they agreed with each survey question. It took less than 30 minutes to complete both closed and open-ended questions. In addition, document analysis was conducted to

review the current training materials. These data were collected during the first stage of the workshop evaluation.

Findings of the pre-workshop survey from raters.

Overall evaluation of the EPT test and rating system. The survey questions about general evaluation of the EPT rating system and the EPT training program in Part 1 provided evidence about how respondents perceived the current EPT test and rating system. Table 1 shows the results of the overal evaluation of the rating system. All respondents answered that they agreed or strongly agreed that the levels of the EPT rating scale (Too low, ESL 113/500, ESL 114/501, and ESL 115/Exempt) were valid for measuring examinees' proficiency levels. The survey showed that the majority of the respondents (87.5%) agreed or strongly agreed that the EPT assessment criteria reflected the ESL writing ability that was being measured. Only 12.5% of the respondents disagreed with this point of view. The respondents positively evaluated the EPT scale descriptors for each proficiency level. For example, 75% of the respondents agreed or strongly agreed that the EPT scale descriptors accurately described each proficiency level. Twenty-five percent of the respondents disagreed with this statement. Based on the results, raters were overall satisfied with the EPT rating system; nevertheless, the findings suggested that the accuracy of the rating scale descriptors and assessment criteria needed to be reviewed.

Table 1

Overall Evaluation of the EPT Test

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
1. I think the levels of the EPT rating scale are valid for measuring examinees' proficiency level (e.g. Too low, ESL 113/500, ESL 114/501, ESL 115/Exempt).	0%	0%	50.0%	50.0%
2. I think the EPT assessment criteria reflect the ESL writing ability that is being measured.	0%	12.5%	50.0%	37.5%
4. Overall, I think the EPT scale descriptors accurately describe each proficiency level.	0%	25.0%	50.0%	25.0%

Findings of the open-ended questions. The open-ended questions were designed to probe into the general evaluation of the EPT essay test. The results were consistent with the findings of Table 1. Six respondents pointed out that as a placement test, the EPT essay test accurately measured writing ability. One respondent wrote:

> The EPT functions well as a placement test. From my teaching experience, there appears to be a good match between students' writing levels and class levels.

Two respondents, however, showed different opinions. This shows that raters were concerned about students' active participation in the test procedures. One respondent wrote:

> In some way, the process is long and it is mostly done early morning. Some students might not give their effort; some might not participate during group activities. However these are mostly things us raters can't control.

This shows that some raters were concerned about the disadvantages of the test topics for individual students. Another respondent wrote:

> Most of the essay tests do. If the question is more difficult to write about for some majors than for others, that might be favor some students even if simply by giving them more time to work on organization- and linguistic- related issues instead of content issues.

Evaluation of EPT test format. In line with the results of Table 1, respondents were asked to evaluate the rating scale descriptors for each proficiency level in Table 2. All participants either agreed or strongly agreed that the EPT scale descriptors for "too low" accurately described examinees' proficiency level. However, participants showed slightly different opinions on the other proficiency levels. Seventy-five of the respondents agreed or strongly agreed that the EPT scale descriptors for "ESL 113/500" accurately described examinees' proficiency level, whereas 12.5% disagreed with the statement. One respondent did not indicate an opinion for this statement. Of the total respondents, 50% agreed that the EPT scale descriptors for "ESL 114/501" accurately described examinees' proficiency level. However, 37.5% of the respondents disagreed with the statement, and one respondent did not respond. Finally, half of the respondents agreed or strongly agreed that the EPT scale descriptors for ESL 115/ Exempt" accurately described examinees' proficiency level. Again, 37.5% disagreed with the statement, and one person did not respond. These findings suggested that the rating scale descriptors for the higher levels were relatively less accurate. In addition, the rating scale descriptors needed to be evaluated and modified through open discussion with raters.

Table 2

Evaluation of EPT Scale Descriptors for Each Proficiency Level

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
5. I think the EPT scale descriptors for "too low" accurately describe examinees' proficiency level.	0%	0%	62.5%	37.5%
6. I think the EPT scale descriptors for "ESL 113/500" accurately describe examinees' proficiency level.	0%	12.5%	62.5%	12.5%
7. I think the EPT scale descriptors for "ESL 114/501" accurately describe examinees' proficiency level.	0%	37.5%	50.0%	0%
8. I think the EPT scale descriptors for ESL 115/ Exempt" accurately describe examinees' proficiency level.	0%	37.5%	37.5%	12.5%

Table 3 showed the evaluation of the assessment criteria. All respondents said that they agreed or strongly agreed that "Organization" reflected the ESL writing ability that was being measured. A total of 87.5% of the respondents agreed or strongly agreed that "Content" and "Grammar and lexical choice" reflected the ESL writing ability that was being measured, but 12.5% of the respondents disagreed with each of these statements. With respect to "use of sources", 75% of the respondents either agreed or strongly agreed that it reflected the ESL writing ability that was being measured. Finally, all respondents agreed or strongly agreed that "Plagiarism" reflected the ESL writing ability that was being measured. These results suggested that raters were confused about content, grammar and lexical choice, and use of sources; however, plagiarism and organization were clearer criteria to raters.

Table 3

Evaluation of Assessment Criteria

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
9. I think "Organization" reflects the ESL writing ability that is being measured.	0%	0%	50.0%	50.0%
10. I think "Content" reflects the ESL writing ability that is being measured.	0%	12.5%	50.0%	37.5%
11. I think "Grammar and lexical choice" reflects the ESL writing ability that is being measured.	0%	12.5%	50.0%	37.5%
12. I think "Use of sources" reflects the ESL writing ability that is being measured.	0%	25.0%	37.5%	37.5%
13. I think "Plagiarism" reflects the ESL writing ability that is being measured.	0%	0%	75.0%	25.0%

Table 4 shows the results of the evaluation of the EPT rating procedures. Based on the results, there was no problem with the double rating system, and all respondents agreed or strongly agreed that at least two raters should score the same essay for the EPT essay test. Only 12.5% of the respondents disagreed that holistic rating method was necessary for the EPT essay test; the same percentage disagreed that analytic rating was necessary. All respondents agreed or strongly agreed that a discussion session with peer raters was helpful for deciding a final score; however, 25% of the respondents agreed that they frequently changed their original score after the consensus process. In addition, they agreed or strongly agreed with a third-rater system being required for solving discrepancies between the decisions of the first two raters. The findings showed that the entire rating system (double rating, holistic rating, and consensus process) was positively evaluated by the raters. One interesting finding was that, although at the time an analytic

rating system was not in use, they felt it could be used in an operational rating situation. Regarding the consensus process, the two raters with the least rating experience, 1 year and 1.6 years respectively, answered that they sometimes changed their final decision after discussion with peers.

Table 4

Evaluation of Rating Procedure

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
14. I think a double rating system is necessary for the EPT essay tests.	0%	0%	25.0%	75.0%
15. I think holistic rating is necessary for the EPT essay test.	0%	12.5%	50.0%	37.5%
16. I think analytic rating is necessary for the EPT essay test.	0%	12.5%	75.0%	12.5%
17. I think a discussion session with peer raters is helpful for deciding a final score.	0%	0%	12.5%	87.5%
18. After the consensus process, I frequently change my original score.	0%	62.5%	25.0%	0%
19. I think a three-rater system is required for solving discrepancies between the decisions of the first two raters.	0%	0%	12.5%	87.5%

Table 5 shows the results of the evaluation of the EPT essay topic. The first three questions asked about the accuracy of the essay topic for measuring the examinees' writing ability. A total of 75% responded that they either agree or strongly agree, suggesting that they felt the three topics accurately measured writing ability. Regarding difficulty, 62.5% responded that they thought the difficulty level of the test topics was different. A total of 37.5 % of the respondents answered that their decisions had been

affected by the topic, whereas the rest (50%) had not changed their decisions because of the topic. In addition, raters agreed or strongly agreed that examinees' writing performance might have differed depending on the topic.

In terms of rater training, 75% of the respondents answered that the training workshop had provided relevant information about the different three topics of the EPT. With respect to the rating process, only 25% of the respondents agreed that they had had some difficulty handling the three different topics when rating. These findings suggest that raters showed a positive attitude toward the accuracy of the test topics for measuring writing ability. The difficulty level of the three topics was not regarded as equivalent. Raters thought overall that the test topic did not influence rating performance, whereas it might have affected examinee performance.

Table 5

Evaluation of the EPT Topic

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
1. I think "Globalization" is a topic accurately measuring examinees' writing ability.	12.5%	0%	37.5%	37.5%
2. I think "Cloning" is a topic accurately measuring examinees' writing ability.	0%	12.5%	37.5%	37.5%
3. I think "Animal Testing" is a topic accurately measuring examinees' writing ability.	0%	12.5%	37.5%	37.5%
4. I think the difficulty level of three different topics is equivalent.	0%	62.5%	12.5%	12.5%
5. I think my final decision is affected by the EPT essay topic given.	25.0%	25.0%	37.5%	0%

(continued)

Table 5 (continued)

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
6. I think examinees' performance is affected by the different topics.	0%	12.5%	50%	25.0%
7. The training materials provided relevant information about the different three topics of the EPT.	12.5%	0%	62.5%	12.5%
8. I think I have a difficulty how to handle three different topics when rating.	37.5%	25.0%	12.5%	12.5%

The survey asked raters about topic difficulty and rating difficulty in Table 6. Four respondents responded that the test topics were equally difficult to write about and rate. T6 answered that topic difficulty was the same under the rating situation. T7 responded that cloning and animal testing were at the same difficulty under testing and rating contexts. As an essay topic, five respondents thought that "Globalization" was the most difficult topic. In line with this, five respondents answered consistently that "Globalization" was also the most intricate topic to rate. These findings suggested that individual raters had different perceptions on the difficulty level of topics, and that the difficulty of topics might have been perceived differently across testing and rating contexts.

Table 6

Different Perceptions on Topic and Rating Difficulty

T	Topic difficulty			Rating difficulty		
	Globalization	Cloning	Animal Testing	Globalization	Cloning	Animal Testing
T1	1	3	2	1	3	2
T2	1	2	3	2	1	3
T3	3	2	1	3	2	1
T4	1	2	3	1	2	3
T5	1	2	3	1	3	2
T6	3	1	1	1	1	1
T7	1	2	2	1	2	2
T8	1	2	3	1	2	2

Note. 1- most difficult, 3-least difficult

Evaluation of workshop program. Table 7 shows the raters' perceptions of how satisfied they were with the current EPT training program. The survey asked about the effectiveness of the training in terms of purpose, organization of the program, their expectations about the program, and the usefulness of the training program. The majority of the respondents (75%) agreed or strongly agreed that high rater reliability was achieved by the EPT rater training; 25% disagreed or strongly disagreed with that statement. Half of the respondents agreed or strongly agreed that high rating accuracy was achieved by the EPT rater training, but 37.5% disagreed or strongly disagreed. A total of 87.5% of the respondents agreed or strongly agreed that the EPT rater training program had been helpful for achieving high agreement between raters. More than 62.5% of the respondents agreed or strongly agreed that the sequence of the program was logically organized, but the rest of them (37.5%) disagreed or strongly disagreed. A total of 62.5 % agreed that the information was effectively presented, but 37.5% of the respondents disagreed or strongly disagreed. Half of the respondents agreed or strongly

agreed that the workshop program met their needs and interests, while the other half disagreed or strongly disagreed. A total of 75% of the respondents felt they were ready to rate essays after the EPT training session, but 50% of the respondents agreed that they needed additional training. These findings suggested that the EPT training focused on rater agreement rather than reliability (consistency) and accuracy. It was suggested that the organization of the workshop program and delivery information be improved. Raters showed self-assurance about rating, although half of them also wanted further training.

Table 7

Evaluation of the EPT Training Program

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
1. High rater reliability has been achieved by the EPT rater training.	12.5%	12.5%	50.0%	25.0%
2. High rating accuracy has been achieved by the EPT rater training.	12.5%	25.0%	37.5%	12.5%
3. High agreement between raters has been achieved by the EPT rater training.	12.5%	0%	62.5%	25.0%
4. The sequence of the program was logically organized.	25.0%	12.5%	37.5%	25.0%
5. The workshop program I attended met my needs and interests.	12.5%	37.5%	25.0%	25.0%
6. I feel I was ready to conduct rate an examinee after the EPT training session.	12.5%	12.5%	37.5%	37.5%
7. The information was effectively presented using visual/audio aids and handouts.	12.5%	25.0%	62.5%	0%
24. I think that additional training is needed for me to rate essays.	25.0%	25.0%	50.0%	0%

With respect to the evaluation of the training materials (see Table 8) used during the workshop, the majority of the respondents agreed or strongly agreed that the training materials had provided relevant information about the purpose of the EPT (75%), about the rating scales (87.5%), and about the consensus process when there was a discrepancy in essay rating (87.5%). In addition, all respondents said that they agreed or strongly agreed that the training materials provided relevant information about the rating procedures for holistic scoring. However, 62.5% of respondents agreed or strongly agreed with the assessment criteria. Regarding the lecture session, 62.5% of the respondents agreed or strongly agreed that the training focus during the lecture session was appropriate for improving rating skills. A total of 37.5 % of the respondents disagreed or strongly disagreed. These findings suggested that the lecture session was not very helpful, and that the training materials needed to be modified to fit the needs of the raters.

Table 8

Evaluation of the Training Materials

Questions	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
8. The training materials provided relevant information about the purpose of the EPT.	0%	25.0%	50.0%	25.0%
9. The training materials provided relevant information about the rating scales.	0%	12.5%	50.0%	37.5%
10. The training materials provided relevant information about the assessment criteria.	0%	37.5%	37.5%	25.0%
11. The training materials provided relevant information about the rating procedures (holistic scoring).	0%	0%	75.0%	25.0%

(continued)

Table 8 (continued)

Questions	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
12. The training materials provided relevant information about the consensus process when there was a discrepancy in essay rating.	0%	12.5%	37.5%	50.0%
13. The training focus during the lecture session was appropriate for improving rating skills.	12.5%	25.0%	50.0%	12.5%

The survey investigated the raters' perception of the prototype sample activity implemented in the workshop session (see Table 9). All respondents agreed or strongly agreed that the prototype samples used during the workshop were appropriate for the training workshop. A total of 75% of the respondents agreed or strongly agreed that the number of prototype samples (12) was sufficient; on the other hand, 25% disagreed. Finally, all respondents agreed or strongly agreed that the prototype samples were helpful to understand how to rate essays. It turned out that prototype sample practice was helpful to understand and improve essay rating, but more practice was suggested.

Table 9

Evaluation of the Prototype Samples

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
14. The prototype samples used during the workshop were appropriate for the training workshop.	0%	0%	75.0%	25.0%
15. The number of prototype samples (12) was sufficient.	0%	25.0%	50.0%	25.0%
16. The prototype samples were helpful to understand how to rate essays.	0%	0%	62.5%	37.5%

Table 10 shows the results of the evaluation of the workshop activities, including individual and group activities, feedback, and the regularity of training session. A total of 75% of the respondents agreed or strongly agreed that both the individual and group rating activities allowed them to acquire practical rating skills. In addition, 75% of the respondents agreed or strongly agreed that the feedback from their peers and the trainer was helpful. A total of 37.5% of the respondents agreed or strongly agreed that the length (3 hours) of the training workshop was appropriate, while 37.5 % disagreed. The majority of the respondents responded that they agreed or strongly agreed that one regular training session per year was sufficient for them to understand the EPT rating system, but 25% disagreed. More than 87.5 % of the respondents agreed or strongly agreed that they would like regular feedback on their rating performance after the EPT; only one respondent disagreed. These findings implied that individual and group activities, as well as feedback, had been useful for improving rating skill.

Table 10

Evaluation of Workshop Activity

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
17. The individual activities during the workshop allowed me to acquire practical rating skills.	0%	25.0%	50.0%	25.0%
18. The group rating activities during the workshop allowed me to acquire practical rating skills.	0%	12.5%	37.5%	37.5%
19. The peer feedback during the group activity was helpful.	0%	12.5%	37.5%	37.5%

(continued)

Table 10 (continued)

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
20. The trainer feedback during the workshop was helpful.	12.5%	12.5%	37.5%	37.5%
21. The length (3 hours) of the training workshop was appropriate for training raters.	0%	37.5%	25.0%	12.5%
22. I think one regular training session per year is sufficient for me to understand the EPT rating system.	0%	25.0%	37.5%	37.5%
23. After the EPT test, I would like to get some regular feedback on my rating performance.	0%	12.5%	62.5%	25.0%

Findings of open-ended questions. The results of the open-ended questions consistently supported the findings. Raters answered that sample practice, group activities, and learning about the test materials (e.g. lecture and reading materials) were useful tools for understanding the rating process and improving their rating skills. Two respondents commented:

> I do remember reading many sample essays and practicing with other students. The extensive pair work that we had was extremely helpful in learning the EPT process and benchmarks.

> What we were given, especially the samples were very good, however, I think it wasn't enough.

With respect to feedback, raters usually received feedback from the trainer (the ESL coordinator) when the encountered difficulty making the final decision. They pointed out that it was helpful to hear contextual justification depending on the different difficulty they were having. In addition, they suggested modifying the workshop program to include more sample practice and feedback. One respondent wrote:

> I usually take the concrete essay to the EPT trainer and ask for feedback on the issue that I am having difficulty with. For instance, if I am not sure that the person has used the sources sufficiently, I take the article and the essay to the EPT trainer, tell her why I am finding it difficult to evaluate that point, and get feedback. I prefer concrete contextualized feedback.

The last question in Part 1 asked respondents to order the workshop activities most to least useful in Table 11. All respondents marked that sample practice was the most helpful. The trainer's feedback was also ranked highly, followed by peer feedback. It turned out that the lecture session was the least helpful.

Table 11

Rank of Workshop Activity

	Workshop Activity				
T	Lecture	Sample practices	Group work	Peers' feedback	Trainer's feedback
T 1	5	1	2	3	4
T 2	5	1	3	4	2
T 3		1			2
T 4	5	1	2	3	4
T 5	3	1	2	5	4
T 6	2	1	3	5	4
T 7		1		3	2
T 8	4	1	5	3	2

Note. (1- most useful, 4- least useful).

Evaluation of the trainer's performance. Part 2 of the survey asked participants to evaluate the trainer's performance in terms of content knowledge, delivery/preparation skills, and responsiveness. More than 80% of the respondents positively evaluated the trainer' performance, indicating that they were overall satisfied with trainer's workshop management ability. However, 12.5% of the respondents had a low opinion of the trainer's content knowledge, responsiveness to individuals when giving feedback, and organization. These findings suggested that these three components should be improved for the next rater workshop (see Table 12).

Table 12

Evaluation of the Trainer's Performance

Question	Percent		
	Poor	Good	Excellent
1. Accurate knowledge of content	12.5%	50.0%	37.5%
2. Delivery skills (lecture, effective examples)	0%	50.0%	50.0%
3. Responsiveness to individuals (feedback)	12.5%	25.0%	62.5%
4. Responsiveness to group (feedback)	0%	25.0%	62.5%
5. Responsiveness to Q & A	0%	37.5%	62.5%
6. Organization of training program	12.5%	50.0%	37.5%
7. Preparation of workshop materials (e.g. sample practice, visual aids, handouts)	0%	50.0%	50.0%

Evaluation of training needs. Table 13 shows that raters evaluated the extent to which they needed further training on rating knowledge, rating skills, and policy. Regarding content knowledge, more than 60% of the respondents answered that further training on the EPT test procedures and the EPT rating scale was not necessary; however, almost half of them agreed that they needed further training about the EPT assessment criteria and rating procedures. In addition, 50% of the respondents answered that some training on the test topics was required.

In terms of rating skills, 62.5% of the respondents agreed that they needed some training or extensive training on internalization of prototype samples for each level and tips on unratable samples. Only 37.5% agreed that they needed further training on rating principles. Most of respondents did not feel to have further training on rating policy.

Table 13

Evaluation of Training Needs

Question	No training required	Need some training	Need extensive training	Not applicable
Content knowledge				
1.Understanding the EPT procedures	62.5%	37.5%	0%	0%
2.EPT rating scale	62.5%	25.0%	12.5%	0%
3.EPT assessment criteria	50.0%	37.5%	12.5%	0%
4.EPT rating procedures	50.0%	50.0%	0%	0%
5.EPT essay topics	37.5%	50%	0%	0%
Rating skills				
1.Principles of rating	50.0%	12.5%	25.0%	12.5%
2.Tips on unratable samples	25.0%	37.5%	25.0%	12.5%
3. Internalization of prototype samples for each level	37.5%	50.0%	12.5%	0%
Knowledge of policy				
1.Essay rating focus	75.0%	25.0%	0%	0%
2.Consensus process	87.5%	12.5%	0%	0%

Findings of open-ended questions. The open-ended questions investigated raters' concerns. As a challenge they might face, raters wrote that they felt confused when essays showed features at different levels across the assessment criteria or when they encountered borderline essays. This indicates that more training or practice might be necessary for internalizing prototype samples across the proficiency levels. Some weakness of the training program that were pointed out were lack of structure, insufficient information about rating materials, and insufficient sample practice. One respondent indicated that absence of teaching experience, tiredness, bad handwriting, and raters making decisions based on personal preference were major sources of difficulty. In addition, it seemed that raters wanted to share knowledge and rating experience, and an extensive training program including a refresher workshop as a continuing education was suggested to resolve their concerns and challenges. Two raters wrote:

> After going through the training process for EPT, I was really concerned of the quality of the raters. The short 30 min training procedure did not really help much in terms of rating. In my case, I relied a lot on my previous rating experience when rating the EPTs. I was grateful I had other experiences in rating essays otherwise; I would have been totally lost.
>
> Longer training programs for inexperienced rater/teachers might be helpful. Much as those raters benefit from the feedback of their more experienced peers, the training may be divided into parts for the novice raters alone, in which they proceed at a slower pace through the practice rating samples and reflect on their choices. The following session, they might work together with the experiences raters to gain perspectives that may come from teaching the writing class. Refresher courses for EPT raters are also a good idea for raters who have not taught or rated in sometime and come back to do that".

Findings of pre-workshop survey from the trainer. The EPT trainer was asked to fill out a pre-workshop survey via e-mail. The survey asked open-ended questions about two topics: the EPT training workshop and workshop preparation/management. The trainer responded that the role of the EPT trainer was to lead the workshop and serve as a fourth reader, providing some justification for ESL placement when disagreement occurred in the rating. Second, she responded that the primary purpose of the EPT rater training is to re-familiarize T.A.s with the EPT essay tests and benchmarks, and to calibrate raters to the benchmarks. Finally, the trainer pointed out that a lack of instructional resources for raters' professional improvement was a big challenge for her. She specifically suggested that instructional materials be standardized and revised to implement a well-organized workshop. This was an interesting answer because raters also addressed this concern in the pre-workshop survey. The trainer wrote:

> The workshop needs to be standardized, and add analytical component….Until now, there really weren't any materials. Development takes time to improve materials implemented this semester.

Part 2 of the trainer's survey asked about plans to organize and manage the rater training workshop. The trainer provided a general idea about the rater training program

based on her rating experience, saying that the workshop program had been designed both to familiarize raters with the EPT rating system and to provide rating practice with prototype essays. Regarding to the focus of the EPT training program, she considered rater reliability and rating accuracy as more important than agreement among raters.

> They don't "have" to agree on placement, but if our test is reliable and rubrics are accurate, the disagreement should come naturally.

She also mentioned her concern that a 2 -3 hour workshop and the limited number of prototype essays were not sufficient to represent the rating issues that EPT raters may encounter: "It's difficult to cover all possible issues with a small set of prototypes, time constraints." These issues were discussed, and some of the content was modified when preparing the new training program.

Findings from the review of the training materials. The ESL course guidelines for internal use, including the prototype essays, were collected and reviewed. In addition, all of the scoring materials provided by the EPT administrator (a graduate assistant) were critically reviewed in order to standardize the training program. The ESL course guidelines provided the EPT benchmarks for both graduate and undergraduate levels along with the different level of ESL writing courses. The holistic benchmarks described each proficiency level and simple directions for connecting essay rating to ESL courses.

The EPT trainer had some old prototype essays and answer keys for rater training. Twelve essays were packaged as a practice set, and an answer key provided the correct proficiency level for each essay, as well as a short justification for the assigned score.

The EPT bulletin provided general issues of the essay test for examinees, including the purpose of the test, test preparation, test process, scoring, and ESL course registration. The EPT administrators are given the most current version of the test

prompts/directions, with test topics and lecture notes that proctors have used in testing situations.

Identifying raters' concerns. The results of the pre-workshop survey showed that raters were satisfied with the EPT rating system overall (double rating, holistic rating, and consensus process, see appendix K); nevertheless, the findings suggested that several area should be improved. First, the findings suggested that the holistic rating scale descriptors needed to be improved. The accuracy of the rating scale descriptors and assessment criteria needed to be reviewed. Particularly, rating scale descriptors at the higher levels (ESL 115, ESL 501, or "exempt") were considered relatively less accurate. In addition, rating scale descriptors needed to be evaluated and modified through open discussion with raters.

Second, in term of test topic and test procedure, it was found that raters showed a positive attitude toward the relevance of the test topics to the measurement of writing ability. However, raters were concerned about the extent of the students' active participation in the test procedures and disadvantages of the test topics for individual students. These findings suggested that individual raters have different perceptions of the difficulty level of the topics, which showed that the difficulty of the topics might affect the rating context. With respect to assessment criteria, some raters felt confused about content, grammar and lexical choice, and use of sources; however, plagiarism and organization had clearer criteria to raters. Interestingly, although an analytic rating system was not used at the time, the raters felt such a system might be useful for reaching a more accurate decision in a practical situation. Regarding the consensus process, the two raters with the least rating experience answered that they were likely to change their

final decisions after discussion with peers. This finding showed that teaching experience was a salient feature when justifying their rating decisions.

In past workshops, the EPT training had focused on rater agreement, rather than on rater reliability (consistency) and rating accuracy. In the new workshop, rating accuracy and consistency needed to be strengthened by revising or creating new rating materials. Next, raters showed enough self-assurance about rating, whereas half also wanted further training. More iterative feedback was suggested for the new workshop, and raters wanted to have more practice rating. In addition, the open-ended questions investigated raters' concerns. Raters pointed out that they felt confused when:

1. essays met the descriptions of different levels across the assessment criteria (e.g. grammar at the ESL 500 level, but organization at the "exempt" level)
2. they had to rate essays on the cusp between two levels.
3. they felt the workshop program was unstructured, the information about the rating materials was insufficient, or the sample practices needed improvement.
4. they had to deal with complicating factors such as the absence of teaching experience, tiredness, bad handwriting, or rating decisions based on personal preference (one respondent indicated that these were the major sources of confusion).

Finally, some problems were clearly identified in terms of the utility of the workshop rating resources. There was no independent EPT workshop package for essay rating, and the training materials were not organized. ESL coursework guidelines provided the only holistic descriptors, and essay rating was not a main issue of the T.A. workshop. In the EPT bulletin, the focus of the information was the description of the EPT test procedures, rather than essay rating.

Increasing accessibility to training materials by synthesizing and reorganizing all relevant materials was an important suggestion. The trainer and the EPT G.A.s had

separate information, and it seemed that they had no opportunity to actively share the rating materials, although the EPT G.A. had access to more information, such as test procedures, test topics, test prompts/directions, and some materials used in the operationalized EPT test. It seemed that the raters had fewer chances to see the rating materials.

The information needed to be updated and revised. Some of the materials had already been updated, but some were out of date. It was pointed out that the prototype essays for practice and their answer keys should be updated, because they no longer reflected the current essay topics and rating issues.

Findings of Research Question 2

Research question 2:

Which part(s) of the EPT training program was (were) standardized and how was the training program standardized?

The second research question was derived from the purpose of Phase 4 (workshop preparation) of the revised Lynch model. To answer the second research question, collaboration with the trainer was necessary. The researcher met with the trainer several times, almost every other week, and made a summary of meeting notes on the results of the discussion with a trainer. The workshop schedule and materials on how to standardize the workshop procedures will be discussed below. All of the discussion about the revision of training materials was uploaded at ESL TA web-blog (http//uiuceslta.blogspot.com/).

Meeting 1. The researcher had the first meeting with the trainer on November 11th, and explained the purposes and schedule of the research. The pre-workshop survey was given to the trainer basic information was shared, such as the characteristics of the

rater group, the nature of the training materials, and the workshop schedule. It was decided that the trainer would review the current training materials and bring some ideas to the next meeting about how to organize the schedule what kinds of training materials to include. In addition, it was agreed that the researcher would provide the results of the pre-workshop survey administered to raters. It took approximately 30 minutes to cover the agenda for the first meeting, and the next meeting was scheduled for November 20th.

Meeting 2. Three main points were discussed at the second meeting to explore the weaknesses of the workshop program. First, the pre-workshop survey results and raters' concerns based on their responses to the open-ended questions were reported: the raters wanted a more organized workshop program with accurate rating scale descriptors, more practice, and more contextualized feedback. Particularly, raters had some difficulty deciding about borderline essays, and they wanted rating principles/tips. The theory from the literature review and the practical findings of the needs analysis from Steps 1 and 2 were used as the basis for revising the existing workshop materials. The following table illustrates the suggestions for the revised workshop organization.

Table 14

Suggestions for New Workshop Program

Area of improvement	Suggestions
Overall organization	The organization of the workshop program and the delivery method of the information should be improved. Training materials should be incorporated and reorganized to fit the EPT rating context.
Familiarization	The lecture session was not very helpful, and it should be more organized. The training materials should be carefully modified to fit the needs of raters and should be updated through stakeholders' evaluation.

(continued)

Table 14 (continued)

Area of improvement	Suggestions
Scoring method	An analytic scoring method can be adopted for the purpose of rating practice. Analytic guidelines provide more accurate information for raters who lack teaching experience. Particularly, in the workshop, raters will use both analytic and holistic scoring methods. New instructional resources should be created and developed to strengthen the current training program.
New essays	New essays should be selected for practice during the workshop.
Sample practice	All respondents consistently marked that sample practice was the most helpful for understanding the rating process, but more practice was suggested.
Activities and feedback	Both individual and group activities were useful for improving their rating skills, including feedback from peers and the trainer. The trainer's feedback was ranked most useful for improving their rating skills.

Meeting 3. Again, rough workshop plans and activities were discussed. First, it was agreed to gather all relevant training materials. Electronic documents containing test directions for examinees, three different test topics, and analytic guidelines were obtained from the EPT G.A. The trainer provided holistic guidelines for the undergraduate and graduate levels, and 12 prototype essay samples used in previous workshops. Finally, the workshop schedule was discussed. The trainer provided a rough idea of the workshop program and wanted to elaborate on her specific plans at the next meeting. The following is the basic outline suggested during this meeting for the workshop program: (a) review of sequence and goals, (b) overview of benchmarks, (c) individual rating & whole group comparison 1 (3 essays), (d) individual rating &whole group comparison 2 (3 essays), (e) discussion of rating procedures (what to do with borderline essays, and what to do when

2 raters disagree), and (f) small group rating and comparison and discussion of trainer feedback.

Employing analytic scoring and selecting new prototype essays were the major issues discussed at this meeting. The researcher proposed that adapting analytic guidelines might be helpful to guide the raters' decision-making processes for borderline or inconsistent essays, because almost half of the raters were new raters with only one semester of teaching experience. It was decided that analytic scoring methods would be employed for the workshop, and an old version of analytic scoring guidelines was modified by the trainer. Regarding to prototype sampling, the trainer wanted to choose new essay samples and revise the answer keys, because they were out of date. The trainer provided the old prototype samples and answer keys for the researcher. In addition, the trainer asked about some issues to be discussed at the next meeting: the principles or standards for prototype essay sampling, and the number of essays to be used for the workshop.

Meeting 4. In early January, the researcher and trainer met again to decide how to conduct the workshop program with the revised training materials, and to confirm the specific workshop schedule. The trainer provided a workshop plan including dates, times, training focus and plans. The workshop was scheduled for January 12th,and the computer lab was reserved for 10:00 a.m. - 1:00 p.m.

The researcher presented a short report for the trainer based on three points, and the trainer's feedback was positive. The researcher double-checked all research plans with the trainer. The researcher asked for some time to have the raters fill out the consent letter before the workshop started, and for permission to audio-record the entire

workshop session. Finally, the researcher asked to collect all of the rating sheets from the activities, and at the end of the workshop, conduct a survey about raters' evaluation of the workshop program. It was agreed that the workshop materials and activities would be developed and modified as the researcher had suggested, but some parts were not accepted due to practical reasons, such as time constraints or fatigue. The trainer developed Power Point slides for the workshop and prepared some activities with the new prototype samples. In addition, the trainer wanted to develop new analytic scoring guidelines and rating sheets for the practice session, rather than modifying the old guidelines. It was decided to have one more meeting before the workshop in order to confirm the workshop materials.

Meeting 5. The fifth meeting was scheduled for January 8th to finalize the workshop procedures and activities. The appropriateness of the scale descriptors, assessment criteria, and scoring methods was evaluated and revised to provide a solution to individual raters' problems. Some parts of the training materials had already been already well prepared, but some had not been because of time constraints.

Meeting 6. The researcher met with the trainer again to double-check the workshop program. This was the last meeting, held on January 11th, one day before the workshop. The trainer shared a web site to which the modified training materials had been uploaded, the UIUC ESL Writing TA home. In addition, the trainer and researcher confirmed the holistic and analytic scoring guidelines, feedback worksheet for obtaining feedback about the revised rubrics from raters, and the rating sheets used for the workshop activities. Figure 11 shows summary of meeting notes below.

Meeting	Agenda	Trainer	Researcher
1	The picture of the new workshop and research schedule were discussed.	Trainer survey was provided.	
2	Analysis of training materials	Trainer provided old training materials for the researcher.	The pre-workshop survey results were reported.
3	Analysis of training materials and sharing outcomes	1) Initial ideas for workshop plans were shared. 2) It was agreed to employ both holistic and analytic scoring methods during the workshop. Trainer revised the analytic scoring guidelines.	Researcher gathered all relevant materials and shared them with the trainer. In addition, researcher made a summary of findings: 1) Researcher reviewed the prototype essays and provided principles for new selections and answer keys so that trainer could choose new essays and make new answer keys. 2) Researcher elaborated on the workshop program based on trainer's initial plans.
4	Modification of workshop materials	Trainer modified the workshop draft, selected prototype samples and rating tips. The workshop program was reorganized, and focus of the workshop was clearly identified (rater reliability, accuracy, agreement).	For the workshop, it was confirmed that researcher should prepare: consent letters for participants, observation notes, audio-recorders, evaluation survey, refreshments.
5	Preparation of the workshop	1) Specific workshop activities (sample practice, trainer feedback, peer feedback,	

Figure 11. Summary of meeting notes.

Figure 11 (continued)

		group work, and lecture) were decided, and trainer made Power Point slides and uploaded the training materials to the web. 2) Prototype samples were decided. 3) Rating package including new analytic scoring guidelines and rating sheet for the practices.	
6	Confirmation of the workshop	Trainer prepared the entire workshop session.	Researcher prepared the research.

Review of old prototype essay. At the meeting, the trainer asked whether the prototype essays were still working for the workshop. The researcher reviewed the 12 prototype essays that the EPT had been using and recommended eliminating some misplaced essays modifying others, which the researcher and trainer discussed. The researcher and trainer agreed that a new selection of prototype essays was essential, because the existing essays were out of date and did not represent the current version of the EPT test. The answer keys were removed, because face to face contextualized feedback was considered more helpful. Standards for essay selection in terms of the number of essays and test topics were suggested. Figure 12 shows the problems of the old prototype essays and suggestions for new ones.

Category	Issues discussed
Prototype Essay	• Provide sufficient materials such as benchmarks (updated version) and reading materials related to the test topic. (These can be used for evaluating source use and citation.) •Prototype sampling for graduate and undergraduate essays should be separated.
Answer Key	• In terms of format, current answer key should be organized using bullet points in terms of overall evaluation, final decision, strengths & weakness (detailed information). • In terms of content, answer key describes general language based on benchmarks, but does not effectively show how the rating scale descriptors would apply to specific cases. • Provide what aspects should be observed, and what evidence should be selected (based on analytic scoring points). • Provide match points/discrepancies between essays and rating scale descriptors. • Provide more specific examples (evidence) for final judgments (e.g. specific sentences, transitions, content). • Provide explanations about the content of a particular essay. • Provide rating principles for borderline essays.
Number of essays	• Level (e.g.113/114/115,500/501/Exempt), topics (globalization/ animal testing/cloning), and status (grad/undergrad) should be considered when deciding the number of essays.

Figure 12. Issues for prototypes.

Figure 12. (continued)

Number of essays	• Prototype 3 x 3 x 2 = 18 (each group has 9 samples) • Borderlines 3 x 3 x 2= 18 (each group has 9 samples) • Each rater group rates 18 essays during practice sessions I and II based on status (grad/undergrad).

Investigation of rating splits from 2009. The researcher reviewed the raters' markings on the EPT essays in order to investigate the rate of disagreement among raters. The purpose of this investigation was to explore what test topics and what proficiency levels frequently influence rater performance, particularly rater agreement. It was expected that the disagreement rate would yield implications for new prototype selection. In the operationalized rating situation, raters normally mark their decision--one holistic score--on the individual essays; therefore, disagreement between two raters could be clearly identified (see Table 15).

The disagreement rate was investigated based on three different topics (See Table 19): "globalization," "animal testing," and "cloning." The rate of disagreement for "globalization" was 18.32%, and it seemed that raters had the most problems distinguishing between the 500 and 501 levels. "Animal testing" had a disagreement rate of 17.53%; the biggest problem with this topic was discriminating between 115 and 114. For "cloning," the rate of disagreement was 33.97%, and raters had the most problems discriminating between both the 500 and 501 levels and the 114 and 115 levels. These findings suggest that there was a larger rating split for the topic of "cloning" compared to other topics, and raters might need more training to rate essays on the border between ESL 114 and ESL115 and between ESL 501 and ESL 500 (see Appendix S and T).

The implication for rater training was that raters should be trained on the cloning topic, and that they need more practice distinguishing between the 114 and115 levels for

undergraduates and the 500 and 501 levels for graduates. Prototype samples for these levels needed to be carefully selected and discussed during the training.

Table 15

Rate of Disagreement Across Three Different Topics

Proficiency level	Rate of Disagreement					
	Animal Testing		Cloning		Globalization	
Ex/501	3	5.88%	0	0%	5	4.72%
501/Exempt	3	5.88%	3	8.11%	6	5.66%
501/500	4	7.84%	6	16.22%	14	13.20%
500/501	8	15.68%	19	51.35%	24	22.64%
500/Exempt	1	1.96%	1	2.70%	0	0%
115/114	13	25.49%	4	10.81%	17	16.03%
114/115	15	29.41%	3	8.11%	27	25.47%
114/113	2	3.92%	1	2.70%	7	6.60%
113/114	2	3.92%	0	0%	6	5.66%
Total	51	17.53%	37	33.97%	106	18.32%

The operational version of the training program.

Standardization of the entire workshop program. This is a general description of the proposed standardized training program. The trainer provided the initial idea for the new workshop program, and the researcher fleshed out the idea. The following is a summary of the changes to the workshop. The workshop content would be presented with Power Point slides; implementation of the practice sessions was discussed. The holistic guidelines were evaluated and the current descriptors were kept analytic scoring guidelines and new assessment criteria were developed. Rater feedback on the criteria and descriptors would be solicited after the workshop. New rating sheets were created for the rating practice. Figure 13 provides the basic outline of the training program.

Stage	Step	Content of workshop
Phase 1	Familiarization	Intro and short lecture session
Phase 2	Norming	Individual & group rating session
Phase 3		Feedback & discussion session
Phase 4	Wrap-up	Closing & evaluation

Figure 13. Workshop program.

Phase 1: Familiarization. The rater training program was modified based on the current content and workshop program of the EPT essay test. The training program was designed to provide short familiarization and norming sessions to understand the EPT rating system. During familiarization, general information on the EPT test and rating procedures would be provided.

First, in the lecture session, knowledge about several rating topics as well as general information would be delivered. The trainer would define the general rating guidelines and provide the raters with information about the purpose of EPT rating, holistic guidelines, test topics, and scoring procedures.

Phase 2: Norming Session: Individual and Group Rating. In the norming session, individual and group practice would be implemented, and trainer feedback would also be provided. In the practice session, raters would practice applying the principles the trainer had provided. Individual practice would proceed first, followed by group discussion for mutual feedback. After the rating, the trainer would confirm the scores of the prototypes with the class, and, if necessary, discuss rating issues again. This process would be repeated until all essay samples were completed. All raters would be involved in the group discussion session to share and resolve their concerns. In addition, raters would be

asked to complete a workshop evaluation survey. The training workshop was expected to last for approximately three hours.

First, in the individual practice session, both holistic and analytic scoring methods would be used. Raters would assign analytic scores first according to the five assessment criteria, and then they would be asked to decide a holistic score–the proficiency level of the essay sample. If they had difficulty determining a score, they would look at the rating scale again and make the best judgment they could. Alternatively, they could make a note of the reasons for their difficulty. The follow-up rater training session would allow them to discuss the issues they had encountered when assigning scores.

Next, in the group practice, raters would make individual decisions first, and then they would discuss the validity of their scores as a group. It was expected that they would also share rating experiences and teaching experiences. They would revise their decisions if necessary, and make notes about what had made them change their scores.

Phase 3: Norming Session: Feedback and Discussion Session. A consensus process using official scores would be the next step. The trainer would provide feedback for the raters, giving the exact scores of the assigned essays, along with an explanation of those scores. The trainer would discuss any discrepancies with the raters. The three different test topics, the diversity of the examinee group, the appropriateness of the operational EPT scoring methods, and the consensus process between two raters would be discussed as well. The rating practice would be repeated until all prototype essays were completed.

Phase 4: Wrap-up. It was agreed that at the end of the workshop, raters would complete a post-workshop survey to evaluate the workshop, and the researcher would

collect the results of the rating process. Figure 14 shows a summary of suggestions for the workshop procedure in detail for each phase.

Step	Organization of workshop program			Suggestions
Phase 1	Familiarization	Lecture	1.Review of sequence and goals	• Review the ESL coursework sequence and goals for each level
			2.Overview of benchmarks	• Review of the EPT benchmarks 1) Holistic guidelines 2) Analytic score guidelines (assessment criteria) • Evaluation of both benchmarks via discussion 3) Review of test topics (three topics)
		Individual & group rating 1	3.Individual ratings & whole group comparison (3 essays)	• Purpose of this section is to help raters internalize prototype samples for each proficiency level. • Essays (prototype samples) from undergraduate level (by considering test topics) will be selected. • Activity procedures: raters do analytic scorings first, and assign a holistic score. 1) Individual rating • Number of essays: 3 prototype essays for each level. • Discussion of how to match essays and descriptors for each proficiency level.
Phase 2	Norming	Individual & group rating 2	4. Discussion, whole group comparison & feedback session	Whole group comparison and trainer feedback
			5. Small group rating and comparison & discussion on trainer feedback (8 essays)	• Purpose of this section is to help raters understand borderline essays and give rating tips about how to handle them. • Essays (prototype samples) from both undergraduate and graduate levels will be selected. • Content of activities, number of essay for each level and procedures are the same as section 3. •Share knowledge, teaching experience, and rating tips.

Figure 14. Suggestions for the EPT workshop program.

Figure 14. (continued)

Phase 3		Individual & group ratings 3	6. Simulation of the EPT rating procedures a. what to do with borderline essays b. what to do when 2 raters disagree	• Purpose of this section is to help raters understand borderline essays and give rating tips on how to handle them. •Essays (borderline essay samples) from both undergraduate and graduate levels will be selected. Activity procedures: raters do analytic scoring first, then holistic scoring. 1) Individual rating 2) Whole group comparison • Number of essays: 2 -3 prototype essays for each level. • Discussion of how to match essays and descriptors for each proficiency level. •Discussion about consensus process when two raters disagree.
Phase 4	Wrap-up	Closing	7. Evaluation of workshop session	•Researcher will provide an evaluation form right after the workshop.

Selection of representative prototype essays. The prototype essays used in the workshop were selected by the EPT trainer. Representative essay samples for varying proficiency levels were selected and used in the practice and discussion sessions of the training workshop in order to enhance raters' understanding of making a scoring judgment. Prototype essay samples were obtained from the EPT data bank (see Table 16 and Appendix M).

Twenty-nine essay samples at both undergraduate and graduate levels were selected, including prototypes of each proficiency level and borderline essays. Three undergraduate prototype essays were assigned for Activity 1. The essays selected for Activity 1 represented the typical characteristics of the three different proficiency levels and were easy to rate. The five analytic scores for these essays supported the holistic score by clearly matching the general features of the essays and scale descriptors. The final scores were provided for raters; Activity 1 was designed so that raters could familiarize themselves with the general features of each proficiency level without guessing.

For Activity 2, eight essays were selected from a pool containing all proficiency levels and both undergraduate and graduate essays. Essays were selected for Activity 2 with the aim of training raters to rate borderline essays. This group of essays included complex features, reflecting inconsistent performance across assessment criteria. These essays required a higher level of cognitive demand or rating strategy, because they did not perfectly match the scale descriptors.

The seventeen remaining essays were assigned for Activity 3, including four undergraduate essays at the 115 level, and six graduate essays at the 500 level. Since

Activity 3 was designed to reflect the experience of rating for the operationalized EPT, essays were randomly selected. The test topic was selected depending on the rating split rate. For practice, an answer key was provided during the feedback session, describing the exact score given by the trainer and the reasoning process for reaching that score. The essay samples used in the post-rating sessions were excluded from the prototype sampling used for training.

Table 16

Prototype Essays Used in the Workshop

	Prototype essays			
Level	Rating scale	Activity 1	Activity 2	Activity 3
Undergraduate	113	3	29	28
	114	1	22, 23	14,15
	115	2	5	12,13,20,21
Graduate	500		7	6,8,19,25,26,27
	501		9, 16	10, 17, 24
	Exempt		8	11
Total		3	8	17

The three test topics were almost equally represented. Ten essays each were written for cloning and animal testing and eight essays were written about globalization. For Activity 1, cloning was selected for all three essays. For Activity 2, three-four essays about animal testing and cloning were selected. Seven-eight essays about animal testing and globalization were selected for Activity 3. These selections were made based on the results of the document analysis: there had been many rating splits for the topic of cloning in the past, so it was assigned in Activities 1 and 2 so that the characteristics of the cloning essay prototypes could be better internalized (see Table 17).

Table 17

Prototype Essay Selection Across the Test Topics

	Essay selection			
Topic	Activity 1	Activity 2	Activity 3	Total
Animal Testing	0	3	7	10
Cloning	3	4	3	10
Globalization	0	1	8	9
Total	3	8	18	

Analytic Scoring Guidelines and Rating Sheets. The EPT trainer developed new analytic scoring guidelines with a 12-point rating scale. In addition, five new assessment criteria were developed by modifying the previous training materials: focus, support/elaboration, organization, conventions, and integration. The first criterion, focus, refers to the degree to which the main idea/theme and point of view are clear and maintained. Support/elaboration indicates the degree to which the main points/elements are elaborated upon and/or explained by specific evidence and detailed reasons. Organization is the degree to which the logical flow of ideas and text plan are clear and connected. Conventions refers to the degree to which the student has mastered grammatical and lexical aspects of English. Finally, integration indicates the overall judgment of how effectively the paper expresses the basic features in order to address the assignment. A rating practice sheet was provided for raters to record both analytic and holistic scores, their opinions, and group activities.

Creating an Interface with Web Tools. All newly developed training materials were uploaded to the UIUC ESL Writing TA home page (http//uiuceslta.blogspot.com/),

including the Power Point slides used in the workshop, the electronic documents about the EPT test procedures, the test directions, test topics, the holistic and analytic scoring guidelines, the rubric revision worksheet for obtaining feedback from raters, and the workshop rating sheet (see Figure 15).

Using a blog is efficient and interactive. EPT raters can access the materials at their convenience without restrictions of time or place. Moreover, raters can leave comments or questions about the rating activities, share opinions with peer raters, and get feedback from the trainer. This could be an ongoing process throughout the semester, and is one way in which the training program could be enhanced for the next workshop.

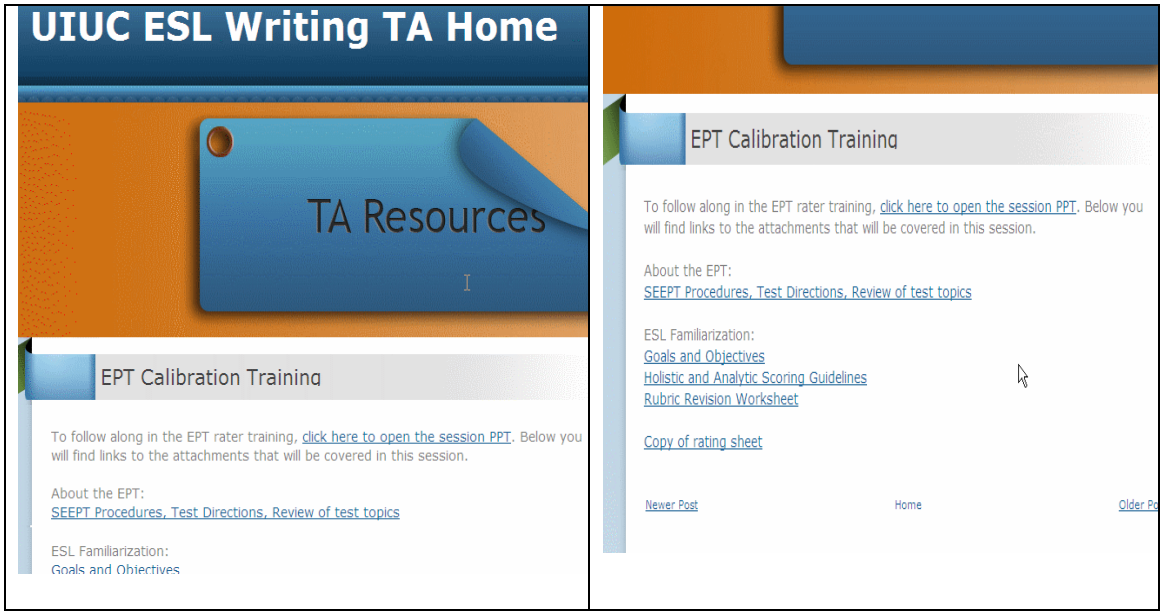


Figure 15. UIUC ESL writing TA homepage on the web.

Findings of Research Question 3

> Was the proposed new training workshop successfully conducted? To identify the qualities of an experienced rater performance through comparison with novice rater performance in terms of attitude, motivation, confidence level, cognitive process, can internalized rules (strategies) on the formulation of a judgment during rating and constructs for expert ratings be identified?

Two workshop sessions were held, on January 12th and 15th, in the Foreign Languages Building (FLB). Through rigorous discussion between the researcher and the trainer based on interim findings from Steps 1, 2 and 3, a clear plan for the revised workshop was formed, and the training workshops were conducted by the EPT trainer.

Workshop participants. Both experienced and newly recruited raters participated in the training. Ten raters participated in the first workshop, and five raters participated in the second workshop. The workshops lasted approximately three hours each. Table 18 shows the background profile of the raters who participated in the workshop.

Table 18

Raters' Profile

	Raters' profile		
Rater	Workshop	ESL Teaching Experience	Rating Experience
R1	1	ESL 500	New
R2	1	ESL 115	New
R3	1	ESL 501,ESL 505	4
R4	1	ESL 113, ESL 500	1
R5	1	ESL 500	New
R6	1	ESL 113,ESL114,ESL115 ESL500,ESL501,ESL505	8
R7	1	ESL114	New
R8	1	ESL114,ESL115	4
R9	1	ESL114	2
R10	1	ESL115	New
R11	2	ESL501	New
R12	2	ESL115	1
R13	2	ESL500	4
R14	2	ESL501	New
R15	2	ESL 115,ESL 501, ESL505	2

Implementation of the workshop.

Familiarization with the EPT rating system. In the lecture session, the EPT essay test and overall rating system were briefly described, including the holistic rating scale, the five assessment criteria, rating methods, and the third rater system. The ESL coursework sequence and the goals for each level described in the ESL coursework handbook were reviewed, and materials related to the EPT rating were also reviewed and discussed, such as holistic and analytic benchmarks and test topics. In addition, the appropriateness of the scale descriptors, assessment criteria, and scoring methods were discussed. Figure 16 is a summary of the revised workshop program. The two workshops followed the same program.

		Purpose	Content
Introduction	Lecture (10:00-10:40)	• to familiarize the EPT rating system	• What is placement? •About the EPT (SEEPT procedures, test directions, review of test topics) •Rating process (goals & objectives, holistic & analytic scoring guidelines, review of test topics) •ESL familiarization (goals & objectives, holistic guidelines, and analytic guidelines) •Read three essay topics •Questions and answers
Familiarization	Activity 1 (-12:00)	• to understand rating process	• Read three essays. Their placement has been provided for you. Consider the placement: is it appropriate? Why? Discuss the benchmarks and scoring guidelines with a partner in order to understand the placement of these two essays. Then, as a group, consider the bench marks, goals and objectives for the courses and see if there are changes that need to be made to the rubric. •Trainer feedback •Reading and rating (do/don't) •Comments and questions
Norming session	Activity 2 (-12:45)	• to calibrate raters	•Read and rate a small set of essays individually (4-5 essays). Then compare your score with the other members of your group. If your scores match, talk about why you chose that placement; if they don't match, come to a consensus about the placement. • Placement problems (consensus procedures, borderline essays, different scores on different criteria, scores below 113 or 500)
	Activity 3 (-1:15)	• do place practice	• As a group, you will be given a stack of essays (2-3 essays). Work together to rate the essays according to the rating procedures; when you are finished, compare your results with the actual placement. If there are discrepancies, discuss why.
Closing	Workshop evaluation (-1:30)		• Evaluation form •Other questions

Figure 16. Revised workshop program.

Focus of activity 1. The purpose of Activity 1 is to familiarize raters with the EPT rating system. Three prototype essays were distributed to raters, and the holistic scores were disclosed. The raters provided analytic scores based on the holistic scores. For this activity, the raters were divided into five pairs. If they had difficulty determining a score, they looked at the rating scale again and made the best judgment they could, or they made a note of the reasons for their difficulty. The follow-up session allowed them to discuss the issues they had encountered.

Findings of activity 1. Table 19 shows comparisons of standard scores with individual rating patterns. The five analytic scores for each essay were added together and converted to a holistic score based on the score range. Raters assigned a holistic score based on global essay evaluation. Four raters' scores across three essays perfectly matched the trainer's scores, but five raters' on Essay 3 were different from the trainer's score. R9 showed some confusion deciding between 114 and 115 as a holistic score on Essay 2, and R12 felt confused between 114 and 115 on Essay 1. It was found that for Essay 3, a holistic score based on global impression was more accurate than the analytic score in comparison with the original and trainer's scores. The holistic decision based on analytic scores was one level lower, and it seemed that the analytic scoring method led raters to assign lower scores. With respect to rating experience, R1, R2, R7, R11 and R14 were new teacher raters who had started to rate essays this spring semester. R3, R8, R9, R13, and R15 had rated essays for at least 2 semesters. The findings suggest that the analytic scoring method should be improved to avoid score lowering.

Table 19

Rating Results of Activity 1

	Rating results of activity 1		
Rater	Essay 1	Essay 2	Essay 3
Original	114	115	113
Trainer	114	115	113
R1	114	115	Too low
R2	114	115	Too low
R3	114	115	Too low
R4	114	115	113
R5	114	115	113
R6	114	115	113
R7	114	115	Too low
R8	114	115	Too low
R9	114	114	113
R10	114	115	113
R11	114	115	113
R12	114	115	113
R13	114	115	113
R14	114	115	Too low
R15	114	115	Too low

Group rating. Raters had one more chance to revise their scores through a discussion with peer raters. Table 20 shows rating accuracy improved. For example, in rating Essay 1, all five groups' scores perfectly matched the trainer's score. For Essay 2, four groups reached a consensus on the ESL 115, but Group 5 considered it to be on the borderline between ESL 114 and 115. Regarding Essay 3, four groups agreed on ESL 113, but Group 1 did reach a consensus. Group 6 in the second workshop showed that the rest of the raters of the second workshop group agreed with the original score after group discussion. The findings showed that group consensus may be helpful for reaching a standard score.

Table 20

Results of the Group Rating

	Results of the group rating					
Essay	G1	G2	G3	G4	G5	G6
Essay 1	114	114	114	114	114	114/115
Essay 2	115	115	115	115	114/115	115
Essay 3	Too low/113	113	113	113	113	113

Training focus of activity 2. The purpose of Activity 2 was to calibrate raters' decision making process. Raters were re-grouped, into groups of five, and each group had the same eight essays. This activity was different from Activity 1 because the individual decision making process was considered more important. No information was provided about the essays, and individual raters spent more time on individual decision making. The focus of the calibration session was how raters apply their understanding of the basic principles of rating. Accurate observation and selection of appropriate information from

the essay were emphasized. Raters obtained feedback from the trainer about how to match essays with scale descriptors for accurate categorization. Through the prototype sample practice, raters could experience evidence-based decision making. Raters assigned analytic scores first, then holistic scores. Otherwise, the rating procedure was the same as Activity 1. Figure 17 shows the training focus of Activity 2.

Rating Process	Training Focus
Accurate Observation	1) How can raters observe the target essays? 2) Holistically? Analytically?
Accurate Selection	1) How can raters select relevant evidence without information loss? 2) Which information is salient for each proficiency level?
Accurate Categorization (Adjustment)	1) How can raters recall correct information from their short term memories? 2) How can raters make correct categorizations? 3) How can raters match the essays with the descriptors?
Judgment	1) How can raters make correct judgments?

Figure 17. Training focus of activity 2.

After finishing the individual rating, raters went through a group consensus process, comparing their individual scores and sharing their experiences and opinions. They discussed holistic or analytic scores with the trainer when a discrepancy arose with the standard scores. Figure 18 shows the salient features distinguishing the proficiency levels to help rate the borderline essays.

	Rating Principles
DO	• Familiarize yourself with the overall scoring procedures. • Understand the test prompts and directions before scoring. • Make scoring judgments according to the rating scale descriptors. • Identify both strengths and weaknesses of the essay. • Think globally instead of picking on isolated errors. • Consider the student's potential for success when determining placement for a borderline essay.

Figure 18. Rating principles proposed.

Figure 18. (continued)

DON'T	• Rely too much on your internal rubric. • Compare one essay to another during rating. • Score an essay without proof or evidence. • Focus only on instance errors. • Score on the factual or other informational content from the sources: reading and lecture. • Focus on bad handwriting or other external factors.

Training focus of activity 3. The purpose of Activity 3 was to simulate and understand the basic principles of the EPT rating situation. As in the EPT rating context, individual raters assigned a holistic score and pairs of raters underwent a consensus process. If there was a discrepancy in scores, they had a group discussion session to discuss the validity of their individual scores. They also shared rating experiences and teaching experiences. After finishing individual and group rating, the trainer confirmed the scores of the prototypes and, if necessary, they discussed rating issues again. For Activity 3, the analytic scoring was excluded. The raters were assigned to groups of five and each group had two-three essays.

Evidence of group consensus process from activity 1, 2 and 3. While the three activities were conducted, the consensus process was audio-recorded and analyzed. The summary of the consensus process yielded some interesting findings. First, raters who presented specific evidence based on the essays and rubrics tended to lead the conversation, and consensus was easily reached. Raters who did not fully understand the meaning of the scale descriptors had trouble rating. In addition, rating experience may have affected the evidence-based decision making process. The more experienced raters tended to present more evidence to support their scores.

Second, raters stated a holistic score or a sum of their analytic scores to initiate the discussion. They preferred to place the proficiency level with the holistic score when the holistic score did not perfectly match with the sum of the analytic scores. In line with this, raters discussed the scores from a global to local perspective; e.g. intro-body-conclusion; paragraph-sentence-vocabulary. Raters who looked at discrete/local evidence tended to have problems making both analytic and holistic decisions.

Third, topic sentences, thesis statements, and organization of ideas were considered important. Writing conventions, such as citation style and handwriting seemed not to affect the final decision. However, raters may have considered the benefits of the ESL course to which the students would be assigned based on their ratings.

Fourth, it is interesting that the more lenient raters focused on the advantages of the essays, while the more severe raters looked at the disadvantages of the essays, citing these features as the main reason for lowering the score.

Fifth, it was found in Activities 2 and 3 in the group discussion (with five raters), that individual raters recognized whether the other raters were harsher or more lenient. It seemed that it might be helpful for individual raters to be aware their own rating pattern, and to adjust their own severity level.

Finally, in terms of rating difficulty and assessment criteria, one rater had difficulty assigning a score when the examinees developed their own ideas without source use or citation. Raters frequently stated that it was somewhat difficult to distinguish between the descriptions of “developing,” “adequate,” and “advanced.” These two statements are related to the need to improve the rating scale and its descriptor.

Closing and evaluation of the workshop. Right after the training, a post-workshop survey was provided for raters in order to explore the degree of their comprehension of the training session and to evaluate its usefulness. The survey took approximately 20 minutes, and it was collected immediately after the workshop program.

Raters' feedback on descriptors. The trainer asked raters to give feedback on the new analytic scoring guidelines after the workshop, and feedback was obtained from the three groups via e-mail. They made comments on three topics: revision of assessment criteria, cut-off scores or analytic score ranges, and the practicality of the analytic scoring method (see Appendix L).

First, they suggested that the descriptors for support/elaboration and organization be revised, because they were ambiguous. For instance, support/elaboration needed additional information to reflect authentic features of essays. Organization included two features: grammar and paragraph development. Raters' concerns arose when the proficiency level of these two features did not correlate. They could be separated as independent criteria. Raters suggested clarifying the organization criterion by adding statements of topic development. In addition, the criterion of integration seemed to be unclear. The following are justification statements from the raters:

"We suggest it includes a mention of the student using his or her own experiences or ideas as support, and not only the sources provided. We found, in one of the essays, that the student used his own ideas to support his claims, but did not cite any of the provided sources. While this is not exactly what we expect for the ESL students, this kind of "support" is not mentioned in the rubric."

"The differences between absent, developing, and adequate aren't very clear, and sometimes we had difficulty judging where to place a student. Even if the grammar was below acceptable, or sources were not cited, at times one student still demonstrated paragraphing (on a low level) – it was just hard to reconcile where to place him."

"As for the integration, it is somewhat vague in terms of analytical evaluation since it is similar to holistic scoring. It would be better if there is an independent descriptor focusing on a thesis statement and topic sentences."

Second, score ranges were frequently mentioned by raters. It was suggested that the scale ranges for each proficiency level be readjusted, because the ranges did not lend themselves to accurate decisions. Raters also suggested adjusting the score range for the developmental stages:

"It also seemed like the score range for 115 was pretty huge; perhaps you could increase the 114 score a little more? Not sure, as the 114 score seemed pretty acceptable."

"It would be better if there is only one score in one descriptor to reduce confusions when rating. It is somewhat hard for raters to decide only two numbers."

"It would be more flexible if descriptors have wider ranges such as 1-5 not 1-2."

Finally, the practicality of using analytic scores will be evaluated by considering time constraints. It should be evaluated whether use of analytic scoring method is worth or worthless for the workshop practice and it should be considered what the benefits of analytic scoring method are.

Findings of Research Question 4

To estimate the effectiveness of the new training program, different measures were conducted, including a post-workshop survey, an investigation of the rating split rate in terms of agreement, and a post-rating discussion session with six raters. The findings of these measures follow.

Research Question 4-1:

Was the proposed training workshop program appropriate for raters?

Participants. A total of 15 raters currently hired as ESL teachers for the spring semester of 2010 participated in both workshops. Eight of the participants were experienced raters with more than 2 semesters of rating experience, while seven of them were new raters with only one semester of teaching experience.

Evaluation of the proposed training workshop. Right after the training, a post-workshop survey was provided for raters to explore the degree of their comprehension of the training session and to evaluate the usefulness of the training session. Table 21 described the results of the overall evaluation of the workshop program. All raters agreed or strongly agreed that the overall training program was well organized, and that the workshop program met their expectations. A total of 93.3% of the raters agreed or strongly agreed that the overall training program was effective for improving their rating skills, and they were satisfied with all aspects of the training materials. There were various opinions about the workshop duration. A total of 66.6 % of the raters agreed or strongly agreed that the training schedule provided sufficient time to cover all of the proposed activities, but 33.4% disagreed or strongly disagreed with this statement. The findings suggested that raters were overall satisfied with the organization of training program, training material/content, and methods, but that a three-hour workshop was insufficient.

Table 21

Overall Evaluation

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
1. The overall training program was well organized.	0%	0%	26.7%	73.3%
2. The overall training program was effective for improving my rating skills.	0%	6.7%	33.3%	60.0%
3. This training program met my expectations.	0%	0%	46.7%	53.3%
4. I was satisfied with all aspects of the training materials.	0%	6.7%	53.3%	40.0%
5. The training schedule provided sufficient time to cover all of the proposed activities.	6.7%	26.7%	26.6%	40.0%

Table 22 shows the results of the evaluation of the workshop program. All raters agreed or strongly agreed with survey items 6 and 8, which indicated that the goals of the training were clearly delivered to raters and that the topics were relevant to raters' responsibilities. Regarding survey item 7, 88.6% of raters agreed or strongly agreed that each session was clearly organized, but 6.7% disagreed with the statement. All raters answered that the lecture and individual practice sessions were helpful to improve their rating, but 88.6 % of the raters agreed or strongly agreed that group discussion was helpful.

All of the raters thought that the prototype essay samples used in the workshop were appropriate, and 93.3% agreed or strongly agreed that the number of prototype samples used in the workshop was sufficient for enhancing their understanding.

Moreover, 93.3% of the raters agreed or strongly agreed that peer feedback was helpful for improving rating.

With respect to time allocation for the workshop, 80% of the raters agreed or strongly agreed that the pace of the training was appropriate, but 20% disagreed with this statement. All raters agreed or strongly agreed that the times allocated for the feedback and discussion sessions were sufficient, and 80% agreed or strongly agreed that the time for the practice session was sufficient to complete all activities.

A total of 93.3% of the raters answered that the delivery methods (power point slides and real EPT essays) were appropriate. All of the raters agreed or strongly agreed that the materials (e.g. holistic and analytic guidelines, rating sheets) were helpful for understanding the content. Most of the raters, 93.3% and 86.7%, respectively, agreed or strongly agreed with usefulness of holistic and analytic scoring methods. These findings indicate that raters were overall satisfied with the new workshop program and its goals, schedule, content and delivery methods, and activities.

Table 22

Evaluation of Workshop Program

Question	Percent				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
6. The goals of the training were clearly defined.	0%	0%	40.0%	60.0%	0%
7. Each session was clearly organized.	0%	6.7%	33.3%	53.3%	6.7%
8. The topics covered were relevant to raters' responsibilities.	0%	0%	13.3%	86.7%	0%

(continued)

Table 22 (continued)

Question	Percent				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
9. The lecture session was helpful for understanding the rating skills.	0%	0%	26.7%	73.3%	0%
10. The individual practice session was helpful.	0%	0%	20.0%	73.3%	6.7%
11. The group discussion was helpful.	0%	6.7%	26.6%	60.0%	6.7%
12. The prototype essay samples used in the workshop were appropriate for enhancing my understanding.	0%	0%	46.7%	53.0%	0%
13. The number of prototype samples used in the workshop was sufficient for enhancing my understanding.	0%	6.7%	46.7%	46.7%	0%
14. The feedback from the peer raters who participated in this training is helpful for improving the quality of my rating.	0%	0%	40.0%	53.3%	6.7%
15. The pace of the training was appropriate for the topics covered.	0%	20.0%	40.0%	40.0%	0%
16. The time for the lecture session was sufficient for understanding the EPT rating system.	0%	6.7%	60.0%	26.7%	6.7%
17. The time for the practice session was sufficient to complete all activities.	0%	20.0%	60.0%	20.0%	0%
18. The times for the feedback and discussion sessions were sufficient for enhancing my rating skills.	0%	0%	60.0%	40.0%	0%

(continued)

Table 22 (continued)

Question	Percent				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
19. Appropriate aids (e.g. audio-visual) for effective delivery were used.	0%	0%	26.7%	66.7%	6.7%
20. The materials (e.g. handouts for activities) provided were helpful for understanding the content.	0%	0%	46.7%	53.3%	0%
21. Analytic scoring method was helpful for improving the quality of my rating.	0%	13.3%	40.0%	46.7%	0%
22. Holistic scoring method was helpful for improving the quality of my rating.	0%	6.7%	33.3%	60.0%	0%

As shown in Table 23, this workshop helped raters learn about EPT rating. A total of 86.7 % of the raters agreed or strongly agreed that this workshop helped them increase their professional knowledge related to essay rating. A total of 93.3% of the raters agreed or strongly agreed that the workshop material were presented at the right level, but 6.7% disagreed with this statement. All raters agreed or strongly agreed that they had internalized the basic concepts of the EPT levels, and 80% agreed or strongly agreed that they had learned problem-solving skills during the discussion. A total of 93.3% agreed or strongly agreed that they had had sufficient opportunity for interactive participation in order to share knowledge and experiences. These findings suggest that the workshop content was helpful for raters, but that they might still want to learn practical rating skills for the operational EPT situation.

Table 23

Evaluation of Learning

Question	Percent				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
23. This workshop helped me increase my professional knowledge (related to essay rating).	0%	6.7%	46.7%	40.0%	6.7%
24. The materials were presented at the right level.	0%	6.7%	26.7%	66.6%	0%
25. I have learned how to internalize the basic concepts of the EPT levels.	0%	0%	60.0%	40.0%	0%
26. I learned problem-solving techniques for essay rating during the discussion session.	0%	20.0%	53.3%	26.7%	0%
27. Sufficient opportunity for interactive participation was provided in order to share different perspectives and experiences with peer raters.	0%	6.7%	73.3%	20.0%	0%

Table 24 shows the findings of the evaluation of application skills. A total of 93.4% of the raters agreed or strongly agreed that the workshop provided balanced content knowledge and practice; however, 6.7% disagreed with this statement. All agreed or strongly agreed that Activity I was helpful for enhancing rating. However, regarding Activities 2 and 3, 6.7% and 13.3% of raters disagreed with the statement. Finally, 86.7% of the raters agreed or strongly agreed that they were able to apply the workshop knowledge to practice situations. These findings suggest that the procedures for Activities 2 and 3 should be reviewed and improved for the next workshop.

Table 24

Evaluation of Application Skills

Question	Percent				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
28. The workshop provided balanced integration between content and practice.	0%	6.7%	46.7%	46.7%	0%
29. Activity I was helpful for enhancing the quality of my essay rating.	0%	0%	53.3%	46.7%	0%
30. Activity II was helpful for enhancing the quality of my essay rating.	0%	6.7%	46.7%	46.7%	0%
31. Activity III was helpful for enhancing the quality of my essay rating.	0%	13.3%	26.7%	60.0%	0%
32. I will be able to put what I have learned in this workshop into practice.	0%	6.7%	40.0%	46.7%	6.7%

Table 25 shows the overall evaluation of motivation level. Three survey items pertained to this category. All raters agreed or strongly agreed that their questions were answered during the workshop. With respect to future workshop participation, 86.7% of the raters agreed or strongly agreed that they would participate in future workshops, but 13.3% disagreed with this statement. Finally, all raters agreed or strongly agreed that they would recommend this workshop to other raters who had not participated in the workshop. These findings suggest that raters felt positively about workshop program.

Table 25

Evaluation of Rater's Motivation

Question	Percent			
	Strongly Disagree	Disagree	Agree	Strongly Agree
33. Most of my questions were answered during the training.	0%	0%	40.0%	60.0%
34. I would definitely participate in a future rater training program.	0%	13.3%	40.0%	46.7%
35. I would recommend this workshop to other EPT raters who did not participate in this workshop.	0%	0%	26.7%	73.3%

Table 26 shows the evaluation of the trainer, whom they evaluated positively. All raters agreed or strongly agreed with all the items for instructor evaluation. These finding suggest that the raters thought the trainer was well organized and prepared the workshop session, was knowledgeable and gave sufficient feedback when raters asked, and used effective delivery methods.

Table 26

Workshop Instructor Evaluation

Question	Percent				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
36. The trainer organized the overall procedures well.	0%	0%	26.7%	73.3%	0%
37. The trainer was knowledgeable about the workshop topic.	0%	0%	20.0%	80.0%	0%
38. The trainer was well prepared for the practice session.	0%	0%	20.0%	80.0%	0%

(continued)

Table 26 (continued)

Question	Percent				
	Strongly Disagree	Disagree	Agree	Strongly Agree	Not Applicable
39. The trainer was well prepared for the discussion session.	0%	0%	26.7%	73.3%	0%
40. The trainer encouraged interactive participation.	0%	0%	20.0%	80.0%	0%
41. The trainer gave sufficient feedback during practice.	0%	0%	26.7%	66.7%	6.7%
42. The trainer clearly answered questions to solve scoring difficulties.	0%	0%	40.0%	60.0%	0%
43. The trainer used effective training methods to deliver the training content and practice sessions.	0%	0%	33.3%	66.7%	0%

Table 27 shows the raters' confidence levels before and after the workshop in terms of content knowledge, ability to apply the content knowledge, and rating scale. First, as to overall content knowledge covered in the workshop, 53.3% of the raters answered that they felt somewhat confident or very confident about the content before the workshop, but their confidence was higher after the workshop: all of the raters felt confident about the content knowledge. Second, 60% of the raters responded that they felt confident about their ability to apply the content knowledge to actual scoring before the workshop. However, after the workshop, all of the raters felt confident about applying the content knowledge. Finally, 66.5% of the raters agreed that they feel confident about their rating skills before the workshop, but 93.3% of the raters felt confident after the

workshop. These findings indicate that the workshop had a positive effect on raters' perceptions about their learning process.

Table 27

Raters' Changes in Perception

Question		Changes in perception			
		Very uncertain	Somewhat uncertain	Somewhat Confident	Very Confident
Overall content knowledge covered in this training program.	Before	13.3%	33.3%	33.3%	20.0%
	After	0%	0%	60.0%	40.0%
Ability to apply content knowledge presented in this training program to actual scoring.	Before	0%	40.0%	46.7%	13.3%
	After	0%	0%	60.0%	40.0%
My rating skills related to scoring.	Before	20.0%	13.3%	53.3%	13.3%
	After	0%	6.7%	46.7%	46.7%
My motivation to be involved with scoring.	Before	0%	40.0%	20.0%	40.0%
	After	0%	6.7%	46.7%	46.7%

Group comparisons of the workshop evaluation. Table 28 shows the results of the mean and standard deviation (SD) for each survey category. The mean and standard deviation were estimated in terms of rating experience and workshop group (first or second workshop). According to the results, the means of the new raters for the categories of overall satisfaction, evaluation of the workshop program, learning, application ability and motivation were relatively higher than those of the experienced raters. Only the mean of instructor evaluation was slightly higher among the experienced raters. This indicates that the workshop was more helpful for new raters than experienced

raters. However, the standard deviations for the new raters were greater than those of the experienced raters.

Ten raters participated in the first workshop (January 12th), and five raters participated in the second workshop (January 15th). Across the six categories, the means of the first workshop group were higher than those of the second group. It is likely that the first workshop was implemented better than the second workshop. A more formal workshop situation seemed to be more helpful to the improvement of rater satisfaction, learning, and motivation. However, one limitation of this analysis is that further analysis to explore whether the group mean differences were statistically significant could not be conducted due to the small number of participants (see Appendix U).

Table 28

Comparison of Group Evaluations

	Experienced			New			First			Second		
Category	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Overall evaluation	8	16.63	2.07	7	17.71	2.75	10	17.60	2.37	5	16.20	2.39
Program	8	56.13	7.28	7	60.29	7.82	10	59.70	6.73	5	54.80	8.87
Learning	8	16.25	2.19	7	16.43	2.23	10	17.30	1.89	5	14.40	.89
Application	8	13.88	2.10	7	13.57	2.37	10	14.30	2.11	5	12.60	1.95
Motivation	8	13.88	1.77	7	14.43	1.90	10	13.90	2.03	5	13.80	1.64
instructor	8	30.25	2.44	7	28.86	3.98	10	30.20	3.33	5	28.40	2.88

Table 29 shows the comparisons of the changes in rater perception by group. The survey asked about three areas of improvement: knowledge, ability to apply their knowledge to practical situations, and rating skill. The results of the group from the first

workshop show that new raters had more changes across all three areas, since the means of the three areas gradually increased. It can be interpreted that new raters in the first workshop felt more confident after the workshop, although their starting point was “very uncertain” or “somewhat uncertain” across the three areas. However, the second workshop presented similar response patterns to the first workshop group in that new raters felt confident in their rating. The results also show that experienced raters’ confidence levels increased after the workshop. This indicates that raters felt that the workshop program was directly helpful and would contribute to their confidence about scoring.

Table 29

Changes in Rater Perception by Group

Area improved		First, Experienced			First, New			Second, Experienced			Second, New		
		N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Knowledge	B	5	3.20	.45	5	1.80	.84	3	3.33	1.12	2	2.00	.00
	A	5	3.60	.55	5	3.20	.45	3	3.67	.58	2	3.00	.00
Application	B	5	3.20	.45	5	2.20	.45	3	3.33	.58	2	2.00	.00
	A	5	3.40	.55	5	3.20	.45	3	4.00	.00	2	3.00	.00
Skill	B	5	3.20	.45	5	1.80	.84	3	3.33	.58	2	2.00	1.41
	A	5	3.40	.55	5	3.00	.71	3	4.00	.58	2	3.50	.71

Notes. B refers to before the workshop, and A refers to after the workshop.

Analysis of open-ended questions. Three open-ended questions were asked to raters. The first question asked about the strength of the current workshop program. Raters pointed out that the calibration session was helpful for understanding rating and improving their rating skills. The revised analytic guidelines were also seen as useful. It was suggested that the assessment components be revised based on the holistic rating scale, and that a numerical rating scale be added in the assessment criteria. It was pointed out that new analytic guidelines might make rating decisions easier. The series of activities with prototype essays, and collaboration with peers were also seen as helpful for raters. Finally, the internet interface for the training contents was also new to them, and they said that they could easily share the updated training materials and their experiences on the web (see Appendix P).

The second question asked about what kinds of content or activities should be added to the training. Most people suggested more detailed explanations and examples for each level, particularly for the analytic scoring guidelines. In addition, one rater stated that a more fundamental understanding of writing itself should be achieved to clearly distinguish among the differences between undergraduates and graduates. It was suggested that some differences between English as a Foreign Language (EFL) and English as a Second Language (ESL) writing contexts be covered as well.

The final question asked raters to give suggestions for future workshop. Raters expressed several different opinions about workshop time management. Some raters suggested having more practice and more detailed explanations of each proficiency level. However, one rater stated that the workshop was longer than expected. Another rater thought that the analytic guidelines were useful, but that the numerical scoring system

should be reconsidered and validated. Another issue was whether holistic and analytic guidelines could be applied to both undergraduates and graduates. Based on rater responses, the findings suggest that there are still issues that need to be improved, although some of the concerns had been resolved.

Rating split 2010 spring semester. After the workshop, 15 raters participated in operationalized EPT rating before the start of the new spring semester. Four EPT tests were administered, and the researcher reviewed the raters' markings on the EPT essays in order to explore the rate of disagreement between raters. The operationalized rating followed the traditional EPT rating system (see Appendix V).

The disagreement rate is one way to estimate the effectiveness of the training program, and it was investigated for three topics: animal testing, cloning, and globalization. Animal testing had a disagreement rate of 16.67%; cloning had a disagreement rate of 22.73%, with the lingering problem of discriminating between the 500 and 501 levels. The rate of disagreement for globalization was 21.92%, and raters had difficulty distinguishing between the 114 and 115 levels. These findings show that there was a larger rating split found for cloning compared to the other topics, and raters might need more training to rate essays on the border between ESL 114 and ESL115 and between ESL 501 and ESL 500. It can be said that training may have had some effect, because the rating split was decreased, but rating problems still remain, as shown in Table 30.

Table 30

Rating Split for Spring EPT Operational Rating

Proficiency Level	Rate of Disagreement					
	Animal Testing			Cloning	Globalization	
Exempt/501	0	0%	1	20%	0	0%
501/Exempt	0	0%	1	20%	0	0%
501/500	1	100%	0	0	2	12.5%
500/501	0	0%	3	60%	1	6.25%
500/Exempt	0	0%	0	0%	0	0%
115/114	0	0%	0	0%	4	25%
114/115	0	0%	0	0%	8	50%
114/113	0	0%	0	0%	1	6.25%
113/114	0	0%	0	0%	0	0%
Total	1	16.67%	5	22.73%	16	21.92%

Research Question 4-2:

> Are there differences in the quality of rating between the workshop group and the control group in terms of both classical (inter- and intra-rater) reliability and rating accuracy as well as a broadened view of it?

Participants in the post-rating session. Six raters participated in the post-rating session. Three of them participated in the spring workshop program, but the rest did not. They all have ESL teaching experience. Three of the raters were experienced raters, and three were new raters, with only one semester teaching experience (see Table 31).

Table 31

Post-rating Participants' Profiles

	Workshop			Control		
Group	R1	R2	R3	R4	R5	R6
Teaching	ESL 114	ESL 500	ESL 113, 500	ESL 114, 115	ESL 501	ESL 114, 115
Rating	New	New	2	3	1	New

Essay selection for post-rating session. The 90 writing samples for the post-rating session were obtained from the EPT bank of approved writing samples, in which names and their school IDs are de-identified. The three topics were evenly chosen and the sample package from various proficiency levels was collected from the EPT essay test to explore the accuracy of the raters' scoring decisions. Thirty essays were chosen for each test topic. Forty-five essays were selected for each examinee group - undergraduates and graduates. Each proficiency level was represented by 15 essays. In addition, perfectly agreed-upon essays and split essays were mixed to explore raters' perception of rating difficulty. The following table shows the distribution of the essays and detailed information about essay selection for the post-rating session (see Figure 19).

Examinee	Level	Essay	Animal Testing	Cloning	Globalization
Undergraduate	113	perfectly agreed-upon essays	2	2	2
		split essays	3	3	3
	114	perfectly agreed-upon essays	2	2	2
		split essays	3	3	3
	115	perfectly agreed-upon essays	2	2	2
		split essays	3	3	3

Figure 19. Essay selection for post-rating session.

Figure 19 (continued)

Graduate	500	perfectly agreed-upon essays	2	2	2
		split essays	3	3	3
	501	perfectly agreed-upon essays	2	2	2
		split essays	3	3	3
	exempt	perfectly agreed-upon essays	2	2	2
		split essays	3	3	3
Total			30	30	30

For the post rating session, 90 essays were carefully selected to gauge rater variability depending on test topic and proficiency level. There was no essay for the "too low" level, because the EPT essay data bank did not contain any essays at that level. For each of the remaining levels, 14 to 16 essays were collected, depending on availability. There were four to six essays for each possible combination of test topic and proficiency level. The following table shows essay information regarding topic and proficiency level.

Table 32

Topic and Proficiency Level Distribution of Essays Used in the Post-rating Session

	Entire set		Animal Testing		Cloning		Globalization	
Level	Frequency	%	Frequency	%	Frequency	%	Frequency	%
113	16	17.8	5	16.7	5	16.7	6	20.0
114	14	15.6	5	16.7	5	16.7	4	13.3
115	15	16.7	5	16.7	5	16.7	5	16.7
500	15	16.7	5	16.7	5	16.7	5	16.7
501	16	17.8	6	20.3	5	16.7	5	16.7
Exempt	14	15.6	4	13.3	5	16.7	5	16.7
Total	90	100	30		30	100	30	100

Raters' performance in terms of severity in the post-rating session. Table 33 shows the severity levels of the six raters, including both the workshop and control groups. Descriptive statistics were employed to describe their rating patterns and calculat e their severity. All six raters scored 90 essays using levels 1 through 4 on the holistic scale, and a range of 3 to 12 for analytic rating. The overall mean of the severity level for holistic rating was 3.00, and the standard deviation was 0.70. For the analytic rating, the mean of the "support/ elaboration" criterion was 7.79, indicating that raters scored harshly across the five criteria. Conversely, raters scored leniently for the "focus" criteria (the mean is 8.24, and SD is 1.54). The means and SD of the holistic scores were almost the same between the workshop group and the control group. For the analytic scoring, the means of the workshop group were slightly higher than those of the control group, but the standard deviation of the workshop group was slightly higher than that of the control group. It is interesting that the workshop group raters scored relatively more leniently than the control group raters.

Table 33

Raters' Rating Patterns in Terms of Severity

Scoring method	Entire Group				Workshop Group				Control Group			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Holistic	1	4	3.00	.70	1	4	2.99	.70	2	4	3.01	.70
Focus	4	12	8.24	1.54	4	12	8.43	1.80	4	12	8.05	1.19
Support	4	12	7.79	1.46	4	12	7.95	1.64	4	12	7.64	1.23
Organization	4	12	8.01	1.55	4	12	8.14	1.76	4	12	7.89	1.29
Conventions	3	12	7.91	1.54	3	12	8.04	1.58	4	12	7.79	1.49
Integration	3	12	8.07	1.53	3	12	8.16	1.68	4	12	7.99	1.35

The means and SD for the holistic scoring were almost the same between the new raters and the experienced raters. For the analytic scoring, the means of new the raters were slightly higher than those of the experienced raters, and the standard deviation of the experienced raters was slightly higher than that of the new raters. For the analytic scoring, the criterion of “convention” was scored more harshly by the new raters (R1, R2, and R6). The score range of the experienced raters (R3, R4 and R5) was wider based on minimum and maximum scores across all scores. The standard deviation supported this statement. These findings suggest that the experienced raters scored relatively more severely than the new raters, and the standard deviation indicates that the experienced raters scored without a central tendency (see Table 34).

Table 34

Raters' Rating Patterns in Terms of Severity

Scoring method	Entire Group				New Rater				Experienced Rater			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Holistic	1	4	3.00	.70	2	4	3.01	.68	1	4	2.98	.71
Focus	4	12	8.24	1.54	4	12	8.60	1.51	4	12	7.88	1.47
Support/ elaboration	4	12	7.79	1.46	4	12	8.14	1.29	4	12	7.45	1.52
Organization	4	12	8.01	1.55	4	12	8.31	1.46	4	12	7.72	1.57
Conventions	3	12	7.91	1.54	5	12	7.86	1.51	3	12	7.97	1.57
Integration	3	12	8.07	1.53	5	12	8.32	1.34	3	12	7.83	1.65

Table 35 shows the severity level of the three raters from the workshop group using descriptive statistics. All three raters scored the same 90 essays. The holistic and analytic scores were analyzed to estimate the raters' severity level for their final judgments. In terms of the use of the rating scale, R1 and R2 used Levels 2 to 4, and R3 used levels 1 through 4. R1 and R2 also displayed a similar rating pattern in using the analytic rating scale, using Levels 6 or 7 on the rating scale. However, R3 tended to use a wider rating scale for both holistic and analytic scores.

When comparing the three raters in the workshop group, the overall severity level of R1, R2, and R3 was similar based on the means of the holistic scores. On the basis of the analytic scores, R1 tended to be relatively more lenient on the focus criterion but harsher on the convention criterion. R2 tended to score more leniently on the focus criterion and more severely on the support criterion. Finally, R3 scored more leniently on the organization criterion, but more strictly on the support/elaboration criterion. R3 tended to score relatively more leniently than the other two raters in assigning analytic scores.

Table 35

Descriptive Statistics of the Workshop Group

Scoring method	R1				R2				R3			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Holistic	2	4	3.09	.66	2	4	3.00	. 69	1	4	3.00	.72
Focus	5	11	8.70	1.56	4	12	9.56	1.45	4	11	9.56	1.39
Support	4	11	8.39	1.57	5	12	8.56	1.29	4	12	8.56	1.53
Organization	4	11	8.43	1.68	4	12	8.98	1.51	4	11	9.98	1.47
Conventions	5	11	8.27	1.46	5	12	8.79	1.30	3	11	9.79	1.46
Integration	5	11	8.49	1.55	5	12	8.97	1.29	3	12	8.97	1.53

Table 36 shows the severity level of the three raters in the control group. All three raters scored 90 essays and used levels 2 through 4 of the holistic rating scale. With respect to use of the analytic rating scale, it turned out that R4 and R5 used a wider rating scale, using levels 5 to 8 across assessment criteria. However, R6 used a more limited range, levels 1- 2, for the analytic scale.

When comparing the three raters, the overall severity level of R4 and R6 was almost the same, but R5 tended to be relatively more lenient then the other two raters. In the analysis of the analytic scores, R6 tended to be relatively more lenient on the convention criterion, but scored severely on the support/elaboration criterion. In R5's rating, the mean of "integration" was highest, indicating that R5 scored most leniently. The convention criterion was relatively harshly scored by R5 and R6. R4 was relatively more lenient than the other two raters in assigning analytic scores, but R6 scored relatively more harshly among the three.

Table 36

Descriptive Statistics of Control Group

Scoring method	R4				R5				R6			
	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD
Holistic	2	4	2.97	.64	2	4	3.10	.75	2	4	2.96	.70
Focus	4	12	8.41	1.46	4	10	8.20	1.19	6	8	7.53	.55
Support	4	11	7.64	1.49	5	12	7.80	1.42	6	8	7.47	.52
Organization	4	12	8.09	1.53	5	11	8.04	1.49	7	8	7.52	.50
Conventions	7	12	9.16	.99	4	11	7.69	1.41	6	8	6.53	.52
Integration	4	12	8.17	1.61	5	12	8.30	1.53	7	8	7.51	.50

Table 37 shows individual raters' descriptive statistics according to test topic, based on the holistic approach and the sum of the analytic scores. In terms of holistic scoring, the means of R3 and R6 for the topic of animal testing were 2.87, indicating that these two raters scored more strictly for the topic of animal testing. For the topic of cloning, the means of R3 and R4 were 2.90, so these two raters scored harshly for this topic. For the topic of globalization, the means of R3 were 2.83, the lowest mean among the six raters. Conversely, R1 and R5 showed consistently higher means among the six raters across the three topics, suggesting that R1 and R5 judge more leniently, while R3 scores more strictly.

In terms of analytic scoring, the means of R3 and R6 for the topic of animal testing were 2.50, the same as their holistic decision making, indicating that these two raters consistently scored the essays harshly for the topic of animal testing. For the topics of cloning and globalization, the means of R3 were 2.40 and 2.37, respectively, indicating that R3 was a tough rater for these topics. Conversely, R1 and R2 showed consistently higher means across the three topics, indicating that there was no difference in rating patterns in terms of severity, according to scoring method. The findings consistently showed that R1 and R2 judged more leniently, and R3 scored more strictly compared to the other raters in both holistic and analytic scoring. These findings suggest that raters score more harshly for the analytic scoring across all three topics.

Table 37

Comparisons of Means and Standard Deviation in Terms of Scoring Methods

Scoring method	Rater	Animal Testing		Cloning		Globalization	
		Mean	SD	Mean	SD	Mean	SD
Holistic	R1	3.10	.66	3.03	.67	3.13	.68
	R2	2.93	.69	3.03	.72	3.03	.67
	R3	2.87	.73	2.90	.76	2.83	.70
	R4	3.07	.58	2.90	.71	2.93	.64
	R5	3.10	.85	3.13	.73	3.07	.69
	R6	2.87	.63	3.10	.66	2.90	.80
Analytic	R1	2.77	.43	2.77	.50	2.87	.51
	R2	2.90	.66	2.80	.48	2.97	.62
	R3	2.50	.57	2.40	.50	2.37	.56
	R4	2.77	.57	2.50	.51	2.67	.66
	R5	2.60	.72	2.60	.62	2.50	.51
	R6	2.50	.51	2.50	.51	2.50	.51

Table 38 shows the comparison between holistic scoring and analytic scoring. The five analytic scores were totaled and converted to a holistic score based on a range. Levels 1-4 were used for the holistic scoring, but only levels 2 to 4 or 3 to 4 were used for the converted holistic scores based on the analytic scores. The means (severity) of the decisions based on analytic scores were consistently more severe across all assessment criteria, and the standard deviation was smaller. It is interesting that the analytic scoring method may have lead raters to rate more harshly.

Table 38

Comparisons of Score Decisions Based on Holistic and Analytic Scores

	Holistic				Holistic based on Analytic scores			
Rater	Min.	Max	Mean	SD	Min.	Max	Mean	SD
Original	2	4	2.99	.81				
R1	2	4	3.09	.66	2	3	2.80	.48
R2	2	4	3.00	.69	2	4	2.89	.59
R3	1	4	2.87	.72	2	4	2.42	.54
R4	2	4	2.97	.64	2	4	2.64	.59
R5	2	4	3.10	.75	2	4	2.57	.62
R6	2	4	2.96	.70	2	3	2.50	.50

Note. N size = 90

Correlation analysis for estimating raters' performance. In terms of individual rater performance, the relationship between holistic scores and analytic scores was identified using the Pearson product-moment correlation (see Table 39). For R1, the corre lation between the holistic and analytic scores showed low association, ranging from 0.27 to 0.37. The criteria of support/elaboration and integration had relatively higher correlations with the holistic score (p = .37). R2's and R3's rating patterns showed a med ium correlation with holistic scores. For R2, the correlation between the holistic and analytic scores ranged from 0.38 to 0.60, indicating a medium association between the tw o. The criteria of support/elaboration and integration were relatively more associated with the holistic score. R3's rating pattern showed that the criteria of support/elaboration and conventions were relatively highly correlated with the holistic score.

The correlation between the holistic and analytic scores of R4 and R5 showed a medium association. In both raters' rating patterns, the criterion of integration was relatively highly correlated with the holistic score. However, the criterion of conventions was relatively the least correlated with the holistic score in R4's rating, and in R5's scoring pattern, the criteria of focus, support/elaboration, and conventions were least related to the holistic score, indicating the same correlation indices. Finally, for R6, the criteria of conventions and integration were relatively highly correlated with the holistic score, but the correlation index for the criterion of organization was not statistically significant.

Table 39

Pearson Product-moment Correlation Between Holistic and Analytic Scores of Raters

	Rater					
Criteria	R1	R2	R3	R4	R5	R6
Holistic	1	1	1	1	1	1
Focus	.32 **	.38 **	.55**	.49**	.47**	.33**
Support	.37 **	.60**	.66**	.49**	.47**	.24*
Organization	.31 **	.44**	.59**	.54**	.64**	.16
Conventions	.27 *	.54**	.66**	.34**	.47**	.52**
Integration	.37**	.57**	.64**	.58**	.62**	.51**

Notes. ** Correlation is statistically significant at the 0.01 level (2-tailed).
* Correlation is statistically significant at the 0.05 level (2-tailed).

Table 40 shows the correlation between the holistic and analytic scores. The second column shows the overall correlation index by all raters' scores. The findings show that the correlation between the holistic score and the criterion of integration was

0.52, which is statistically significant at the .01 level, indicating that association was relatively higher than for the other assessment criteria. In the workshop group, the criteria of support/elaboration and integration showed significantly medium association with the holistic score, and in the control group, the criterion of integration was relatively highly associated with the holistic score, which showed medium association. Regarding rating/ teaching experience, the experienced group showed a medium association with the holistic score, but the new rater group showed a low association with the holistic score. These findings suggest that the correlation overall exhibited a medium-low association between holistic and analytic scores, which was significant at the .01 level. In addition, the findings did not show a halo effect, since the existence of halo effect is indicated by high correlation (r. ≥8).

Table 40

Comparisons of Correlation Analysis of Different Groups

	Group				
Criteria	All	Workshop Group	Control Group	Experienced Group	New Group
Holistic	1	1	1	1	1
Focus	.39**	.39 **	.41**	.49**	.29**
Support	.47**	.53**	.42**	.55**	.39**
Organization	.45**	.43**	.48**	.59**	.30**
Conventions	.38**	.48**	.29**	.43**	.33**
Integration	.52**	.51**	.54**	.61**	.40**

Notes. ** Correlation is statistically significant at the 0.01 level (2-tailed).
* Correlation is statistically significant at the 0.05 level (2-tailed).

Rating accuracy. Rating accuracy was analyzed and estimated by directly comparing the raw scores of the originals with the raters' scores. Table 41 shows the level of exact agreement between the holistic scores of the six individual raters and the or iginal holistic scores. The average percent of agreement for all six raters was 48.70 %, a low agreement rate. R1 and R2 had a 53% occurrence of perfect agreement and R3 and R4 reached 42% and 48 %, respectively. R5 had 53%, and R6 had 42 % agreement (see Appendix W and X).

The analytic scores of the raters were analyzed by totaling the five analytic scores. The trainer provided a score range for recalculating an analytic-based holistic score. The purpose of this analysis was to evaluate the relationship between an impression-based holistic score and an analytic-based holistic score. The accuracy of the analytic scoring was slightly lower than that of the holistic scoring, when the scores were compared to the original scores. Table 41 shows rating accuracy depending on scoring method.

Table 41

Comparisons of Rating Accuracy With Respect to Scoring Method

	Holistic scoring		Holistic based on analytic scoring[6]	
Rater	Frequency	Percent	Frequency	Percent
R1	48	53%	39	43%
R2	48	53%	42	47%
R3	38	42%	32	36%
R4	43	48%	36	40%

(continued)

[6] For undergraduates, 1-15 : too low, 16-30:ESL 113, 31-46:ESL 114, and 47-72: ESL 115. For graduates, 1-18: too low, 19-46: ESL 500, 39-46: 47-57, and EXEMPT: 58-60.

Table 41 (continued)

Rater	Holistic scoring		Holistic based on analytic scoring	
	Frequency	Percent	Frequency	Percent
R5	48	53%	33	37%
R6	38	42%	29	32%
Total	263/540	48.70%	211	39.07%

Table 42 shows some interesting results. Both the workshop group and the control group showed more accuracy in the holistic scoring than the analytic scoring. An interesting note is that in group comparisons, the workshop group scored more accurately, but the new raters rated more accurately than the experienced raters.

Table 42

Group Comparisons of Rating Accuracy

Group	Holistic scoring		Analytic scoring	
	Frequency	Percent	Frequency	Percent
Workshop	134/270	49.62	113/270	41.85
Control	129/270	47.78	98/270	36.29
Experienced	129/270	47.78	101/270	37.41
New	134/270	49.62	110/270	40.74

Table 43 shows rating accuracy across the three test topics. For holistic rating, rating accuracy for the topic of globalization was the highest, but accuracy was low for the topic of animal testing. For analytic scoring, the rating accuracy for the topic of animal testing was the highest, but the rating accuracy for the topic of globalization was the lowest. This inverse result indicates that scoring method might affect rating

performance.

It is interesting that rating accuracy might be variable depending on the scoring method across test topics. In addition, when the analytic scoring method was used, raters scored more accurately for the lower levels, such as ESL 113 or ESL 500, and for undergraduate students. Conversely, it seemed that the holistic scoring method was more effective for scoring the higher proficiency level essays and graduate student essays, although there was variability among raters (see Appendix X).

Table 43

Rating Accuracy Across the Test Topics

		Rater							
Scoring Method	Topic	R1	R2	R3	R4	R5	R6	Total	Percent
Holistic	Animal Testing	18	17	10	15	16	13	89	49.44%
	Cloning	18	15	13	17	19	10	92	51.11%
	Globalization	12	16	15	11	13	15	82	54.56%
Analytic	Animal Testing	14	16	10	15	11	10	76	42.22%
	Cloning	12	14	12	8	13	10	69	38.33%
	Globalization	13	12	10	13	9	9	66	36.67%

Cohen's kappa coefficient for agreement. Cohen's Kappa coefficient is used to a measure the reliability between two raters (intra-rater reliability), and it can provide statistically more meaningful estimation than simple percentage of agreement. The value of Kappa can be calculated between 0 and 1, and two raters are in perfect agreement when the value is 1. For this study, the kappa coefficient was estimated between the original scores and the scores of the six individual raters. According to Table 44, R1, R2, and R5 showed a fair level of agreement with the original holistic scores, while the remaining three raters showed a low level of agreement with the original holistic scores.

Overall, all six raters showed a low level of agreement, which supports the findings of the simple percentages of agreement.

Table 44

Agreement Between Original Score and Individuals Using Cohen's Kappa

	Agreement		
Rater	Kappa Coefficient	Asymp Standard Error	Approx. Sig.
R1	.301	.076	.000
R2	.300	.078	.000
R3	.138	.077	.056
R4	.216	.078	.002
R5	.301	.078	.000
R6	.133	.076	.065

Rater reliability.

Analysis of G-study. Based on the statistical analysis, a G-study was carried out t o see the relative variance for each facet. Table 45 shows the results of the G-study (P x R design) for the holistic scores of the six raters. The results indicate that the variance co mponent for person effect was 0.2215 with 0.0394 standard error, which accounted for 45.31% of the total variance. The variance component for rater effect was 0.0048 with 0.0041 standard error, which corresponds to 0.99% of the total variance. A total of 53.71% of variance came from the interaction effect between person and raters aspects. These findings indicate that the rater effect was minor and major source of variance was examinees and interaction effects.

Table 45

Variance Components for Source Effect: [P x R Random Design]

Effect	Variance components		
	Variance	%	SE
P	.2215	45.31	.0394
R	.0048	.99	.0041
PR	.2626	53.71	.0175
Total	.4889	100	

Table 46 shows the results of the G-study for the analytic scores of the six raters in terms of assessment criteria. The results indicate that the variance component for person effect was 0.3208 with 0.1345 standard error, which explained 2.94% of the total variance. The variance component for the assessment criteria effect was 0.2018 with 0.2001 standard error, which corresponds to 1.84% of the total variance. There was no rater effect found, as shown in Table 58. It can be noted that variance in performance can be explained largely by the interaction of the essay/raters with the criteria as well as the interaction among the examinees, raters, and criteria. The interaction effect among these three facets contributed to 95.22% of the total variance.

It is interesting to note that the large amount of interaction between raters and assessment criteria and interaction constituted almost 95% of the total variance. This suggests that the way the essay/rater interacts with the assessment criteria can affect the scoring performance. Since rater effect was not detected for this study, further analysis, such as group differences using G-study, was not conducted.

Table 46

Variance Components for Source Effect: [P x A x R Random Design]

Effect	Variance components		
	Variance	%	SE
P	.3208	2.94	.1345
A	.2018	1.84	.2001
R	0.0	0	.0587
PA	1.1334	10.34	.1961
PR	.1607	1.47	.1366
AR	. 6614	6.03	.2278
PAR	8.4826	77.39	.2841
Total	10.9611	100	

FACETS analysis. Figures show the FACETS summary map for both holistic and analytic scoring for the EPT raters, essay, and test topics (See Appendix Y and Z). The first column shows the severity variation among raters. The most severe rater was at the top, and the least severe rater was at the bottom. The essay column shows that the essay on the EPT demonstrated a wide range of abilities. The spread of the essay ranged from -4 to +4 on the logit scale. The essays with high scores are at the top of the scale, whereas the lower scores are at the bottom. The third column presents the difficulty of test topics. The most difficult topic is at the top and the easiest topic is at the bottom, logit scale 0. In the last column, the intervals of the rating scale used in the study are displayed. A more detailed description of severity and consistency for each facet is reported in Table below. In addition, Appendix Y is a summary map for the analytic scoring method; as for

the criteria in the rating scale, these were quite separated in their place on the logit scale. The criteria above 0 on the logit scale were considered to be more difficult.

Table 47 presents measures of rater performance in terms of both holistic and analytic scoring methods. The holistic score and analytic scores were estimated differently because the rating scale system was different. The holistic scale has four levels, but the analytic scale has 12 levels. The six raters who participated in this study ranged from -3.11 to -2.09 in severity for the holistic scoring. For analytic scoring, rater severity ranged from -1 to 0.7 on the logit scale. On average, the raters below or closer to 0 on the logit scale were more lenient. For the holistic rating, the toughest rater was Rater 3 (-2.09), and the most lenient was Rater 1 (-3.11). For the analytic scoring, R3 (0.71) was still the toughest, and R2 (-1.00) was the most lenient.

Table 47 shows the fit values for the different raters. Three standards can be applied for deciding a misfit rater. First, if the infit value ranges from 0.5 to 1.5 (Lunz et all., 1990), it indicates that the rater scored the essays consistently. Based on this standard, there were no misfit raters found in this analysis. If we use the recommended standard, which is [1.00 ± (.20 x 2 = .40)] for the holistic scoring method, and [1.00± (.19 x 2 = .38)], then no raters were found to be misfits, as they were within the acceptable ranges of 0.60 to 1.4 or 0.62 to 1.38, which means that all raters were consistent in their own ratings across the different scoring methods.

However, if we adopt a more conservative measure recommended by McNamara (1996) for misfits [below 0.8 or above 1.2], in the holistic rating, Raters 1and 3 would be considered inconsistent in their own ratings. Rater 1 might show an overfit rating pattern by using a limited range of the rating scale, while Rater 3 showed a misfit rating pattern,

or inconsistent rating. For the analytic method, Raters 3 and 4 would be regarded as inconsistent, and Raters 5 and 6 showed overfit rating patterns with boundary values below or around 0.8. It is interesting that Rater 6 showed different rating patterns across scoring methods; for the holistic scoring, R6 showed inconsistent rating but for the analytic method, R6 showed an overfit rating pattern.

With respect to agreement calculated with Rasch analysis, there was a moderate 52.5% agreement among the raters for the holistic method, and a low agreement of 20.6% for the analytic method. The separation indexes were 1.29 and 13.26, and the Chi-square value was significant, implying that the raters in this study differed from one another in their ratings (Linacre, 1989).

Table 47

Comparisons of Analysis of Intra-rater Reliability

	Holistic method					Analytic method				
Rater	Severity (Logit)	SE	Infit MS	Z	Exact Agree %	Severity (Logit)	SE	Infit MS	Z	Exact Agree%
R1	-3.11	.23	.78	-1.7	54.7	-.57	.04	.96	-.5	22.5
R2	-2.70	.23	1.05	.3	52.4	-1.00	.04	.95	-.7	18.8
R3	-2.09	.23	1.14	.9	48.9	.71	.04	1.26	3.5	16.1
R4	-2.55	.23	.82	-1.3	56.2	-.43	.04	1.23	3.2	23.8
R5	-3.16	.23	.85	-1.1	55.1	-.18	.04	.75	-4.0	22.8
R6	-2.50	.23	1.35	2.3	47.8	.43	.04	.82	-2.7	19.6
M	-2.69	.23	1.00	-.1		-.17	.04	1.00	-.2	
SD	.37	.00	.20	1.4	52.5	.58	.00	.19	2.8	20.6

Note. Holistic: Separation: 1.29, reliability .62, Chi:15.9, df=5, sig=.01
Analytic: Separation: 13.26 (14.53), reliability:.99(1.00) chi:1056.7, df=5, sig=.00

Table 48 shows the difficulty of the test topics in relation to the scoring method. The topic of animal testing (-0.04) was relatively leniently scored, while the topic of globalization was relatively strictly scored for the holistic scoring. For the analytic scoring, the raters seem to be more lenient with the topic of animal testing (-0.03), but they were stricter on the topic of cloning. None of the test topics were found to be misfits, as they were all within the acceptable range of 0.94 to 1.06, [1.00 ± (0.03 x 2 = 0.06)] for the holistic method, and within the range of 0.8 to 1.2, [1.00 ± (0.10 x 2 = 0.20)] for the analytic method. This implies that the test topics were scored consistently, with a similar severity level.

Table 48

Comparisons of Analysis of Test Topics

	Holistic scoring				Analytic scoring			
Test Topic	Severity (Logit)	Model S.E.	Infit MS	Z	Severity (Logit)	Model S.E.	Infit MS	Z
Animal Testing	-.04	.16	1.02	.2	-.03	.03	1.13	2.7
Cloning	-.01	.17	1.02	.2	.04	.03	.89	-2.3
Globalization	.05	.16	.95	-.4	-.01	.03	.96	-.8
M	.00	.16	1.00	.0	.00	.03	1.00	-.2
SD	.04	.00	.03	.3	.03	.00	.10	2.1

Table 49 shows information about the five assessment criteria. In terms of difficulty, we can see that the raters seemed to be more lenient with the criterion of integration (-0.32) but stricter on the criterion of support/elaboration (0.35). In terms of infit value and Z-score, none of the criteria were found to be misfits, since they were all within the acceptable range of 0.84 to 1.16, [1.00 ± (.08 x 2 = .16)]. This implies that the

five assessment criteria were consistently scored, and they might be considered appropriate criteria for assessing the EPT essays.

Table 49

Analysis of Consistency of Assessment Criteria

	Fit Statistics			
Assessment Criteria	Severity	Model S.E.	Infit MS	Z
Focus	.02	.04	.89	-1.8
Support/elaboration	.35	.04	1.06	.9
Organization	.17	.04	.99	.0
Convention	-.23	.04	1.11	1.7
Integration	-.32	.04	.94	-1.0
M	.00	.04	1.00	-.1
SD	.25	.00	.08	1.3

Interaction (bias) analysis. Analyses of the interaction between raters and test topics and between raters and assessment criteria were conducted using FACETS. Bias can be detected when raters' severity level was gradually more lenient or harsh. No interaction effect between raters and test topic was found for holistic analysis. However, for analytic scoring, three statistically significant cases were found (see Table 50).

The statistically meaningful bias sizes were 0.17, 0.18, and .019 logit. For instance, R4 seemed to score the topic of animal testing more severely than R4's general scoring pattern. R1 and R2 seemed to rate the topic of globalization more harshly by 0.18 logit or 0.19 logit. With respect to the interaction of raters and assessment criteria for the analytic scoring method, seven interaction effects were detected. R2 showed a bias on the criterion of focus (0.30 logit). R4 showed a bias on the criteria of support/elaboration (-

0.38 logit) and conventions (0.81 logit). R4 severely scored 1.19 logit (absolute value) with conventions than with support/elaboration. R5 also showed interaction with the criteria of conventions (-0.20) and integration (0.20). R6 showed bias for conventions (-0.62) and support/elaboration (0.32). These findings suggest that all of the control group raters, except R2, showed interaction with the assessment criteria. This is useful information for future training programs, because the suggested criteria can be re-analyzed and modified. In addition, raters should be re-trained on the test topics in order to remove any interaction effects (Linacre, 1989).

Table 50

Results of Interaction Effects

	Fit statistics							
Interaction	Obs-Exp Average	Bias Size	Model S.E.	Rater	Test Topics	Obsvd Count	t	Sig.
Rater by	.20	.17	.07	R4	animal testing	150	2.25	.027
Topic	.22	.18	.07	R1	globalization	150	2.47	.015
	.23	.19	.07	R2	globalization	150	2.56	.011
Rater by	-.69	-.62	.10	R6	convention	90	-6.21	.000
Assessment	-.43	-.38	.10	R4	support	90	-3.83	.000
Criteria	-.22	-.20	.10	R5	convention	90	-2.00	.048
	.23	.20	.10	R5	integration	90	2.01	.047
	.35	.30	.10	R2	focus	90	3.04	.003
	.36	.32	.10	R6	support	90	3.24	.002
	.96	.81	.10	R4	convention	90	8.39	.000

Analysis of essays used in the post-rating session.

Fit statistics of essays. Ninety essays were analyzed using FACETS, and their fit statistics were estimated. The standard range of 0.5 – 1.5 was applied for the fit statistics, because the standard deviation estimate for fit range was too wide[7], and the range of 0.8 – 1.2 was too conservative. Below 0.5 was considered overfit, and more than 1.5 was

[7] For holistic scoring, fit range was [1.00 ± (.52 x 2 = 1.04)] = 0 – 2.04.
For analytic scoring, fit range was [1.00 ± (.55 x 2 = 1.1)] = 0- 2.1.

regarded as mistfit. Table 51 shows the fit statistics for the essays used in the post rating session. In terms of scoring method, more misfit essays were found for the holistic rating; however, for the analytic scoring, the number of overfit essays was almost equivalent to misfit essays. Regarding test topic, essays with the topic of cloning showed overfit or misfit, and the undergraduate level essays showed misfit compared to the graduate level essays (See Appendix Z). Essays 57, 62, and 72 showed overfit for both scoring methods, and Essays 3, 5, 19, 21, 24, 48, and 84 were considered misfit for both scoring methods. These findings suggest that the characteristics of these essays should be re-analyzed as to what aspects led raters to exhibit misfit/overfit rating patterns for these essays. This information can be used for the next workshop preparation (See Appendix AA).

Table 51

Fit Analysis of Essays Used in the Post- rating

		Holistic Method		Analytic Method	
Examinee	Test Topic	Overfit	Misfit	Overfit	Misfit
Undergraduates	Animal Testing		3, 5, 10		3, 4, 5, 6, 11, 14
	Cloning		19, 21, 24, 26	20, 23, 25, 30	18, 19, 21
	Globalization	36	33	42, 43	37
Graduates	Animal Testing	57	48	57	46, 48
	Cloning	62, 72,75	63, 64	62, 67, 69, 72	61
	Globalization		84, 90	77, 81	84
Total		5	13	14	13

Rater perception of difficulty level. Difficulty level was determined based on the pre-assigned scores. For example, a difficulty level of 1 was assigned when two raters had perfectly agreed on the holistic score. Level 2 was assigned when two raters disagreed but one rater showed confusion between the right proficiency level and one level higher or lower; for example, if the first rater scored the essay as 113, but the second rater scored it as 113 or 114. It is likely that Rater 2 was confused between 113 and 114, and this essay would be considered Level 2 difficulty. Level 3 was assigned when two raters disagreed and a third rater was involved (See Appendix BB).

Table 52 shows the difficulty of the selected essays used in the post-rating. It was determined that 43.3% of the essays were considered easy, 18.9% were medium difficulty, and 37.8% were regarded as difficult. The following table also includes the distribution of difficulty by test topic.

Table 52

Difficulty of Essays Used in the Post-rating

	All Essays		Animal Testing		Cloning		Globalization	
Difficulty	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Easy	39	43.3%	14	46.7%	11	36.7%	14	46.7 %
Medium	17	18.9%	4	13.3%	7	23.3%	6	20.0%
Difficult	34	37.8%	12	40.0%	12	40.0%	10	33.3%
Total	90	100	30	100	30	100	30	100

In the post rating session, raters were asked their perception of the difficulty level when rating (see Table 53). Raters' perceptions were compared to the rating difficulty as assigned above. R1 responded that 42.2% of the essays were easy to rate, 38.9% were medium difficulty, and 18.9% were difficult. The responses of R2 and R3 were similar in that they thought the majority of essays were easy; 82.2% and 91.1%, respectively, 16.7% and 8.9% were medium difficulty, and no essays were difficult. R4 answered that 77.8% of the essays were medium difficulty and 22.2% were easy or difficult. R5 and R6 reported that all essays were medium level. R1's perception was the most similar to the originally assigned difficulty levels; the rest of the raters showed different perceptions of rating difficulty.

Table 53

Individual Rater Perceptions of Essay Difficulty

	Easy		Medium		Difficult	
Rater	Frequency	Percent	Frequency	Percent	Frequency	Percent
Original	39	43.3%	17	18.9%	34	37.8%
R1	38	42.2%	35	38.9%	17	18.9%
R2	74	82.2%	15	16.7%	0	0%
R3	82	91.1%	8	8.9%	0	0%
R4	12	13.3%	70	77.8%	7	7.8%
R5	0	0%	90	100%	0	0%
R6	0	0%	90	100%	0	0%

As further analysis, rating difficulty was re-analyzed by test topic, as shown in Table 54. For the topic of animal testing, R1 responded that 83.3% of the essays were

either easy or medium difficulty. R2 and R3 answered that most of the essays were easy, but R4 showed different perceptions of rating difficulty, reporting that 20% of the essays were easy. R5 and R6 marked that all essays were medium difficulty. For the topics of cloning and globalization, the raters showed similar patterns in perception of rating difficulty. These findings show that raters consider the essays easy-medium to rate, and that raters scored all of the essays without much difficulty.

Table 54

Rating Difficulty by Test Topic

		Easy		Medium		Difficult	
Test topic	Rater	Count	Percent	Count	Percent	Count	Percent
Animal	R1	13	43.3%	12	40.0%	5	16.7%
	R2	27	90.0%	2	6.7%	0	0%
Testing	R3	26	86.7%	4	13.3%	0	0%
	R4	6	20.0%	21	70.0%	3	10.0%
	R5	0	0%	30	100%	0	0%
	R6	0	0%	30	100%	0	0%
Cloning	R1	13	43.3%	12	40.0%	5	16.7%
	R2	25	83.3%	5	16.7%	0	0%
	R3	29	96.7%	1	3.3%	0	0%
	R4	3	10.0%	26	86.7%	1	3.3%
	R5	0	0%	30	100%	0	0%
	R6	0	0%	30	100%	0	0%
Globalization	R1	12	40.0%	11	36.7%	7	23.3%
	R2	22	73.3%	8	26.7%	0	0%
	R3	27	90.0%	3	10.0%	0	0%
	R4	3	10.0%	23	76.7%	3	10.0%
	R5	0	0%	30	100%	0	0%
	R6	0	0%	30	100%	0	0%

Table 55 shows the comparisons of the fit statistics and rating difficulty. Fit statistics were measured with FACETS and compared with the original rating difficulty. A total of 47% of the total essay which showed misfit/overfit rating patterns were

perceived as difficult. A total of 61% of the overfit essays were marked as difficult, and 37 % of the misfit essays were seen by raters as difficult. These findings indicate that rating difficulty may be associated with inconsistent rating patterns.

Table 55

Comparisons of Fit Statistics and Rating Difficulty

Holistic Scoring Method				Analytic Scoring Method			
Overfit		Misfit		Overfit		Misfit	
Essay	Difficulty	Essay	Difficulty	Essay	Difficulty	Essay	Difficulty
36	3	3	1	20	1	3	1
57	3	5	3	23	3	4	3
62	1	10	1	25	3	5	3
72	3	19	3	30	3	6	3
75	2	21	3	42	1	11	3
		24	1	43	1	14	2
		26	1	57	3	18	3
		33	2	62	1	19	3
		48	1	67	3	21	3
		63	1	69	3	37	1
		64	1	72	3	46	1
		84	1	77	3	48	1
		90	1	81	1	61	1
						84	1

Analysis of raters' reflection logs. Raters who participated in the post-rating session created a written reflection log after finishing their essay rating. They made statements about variety topics, including the holistic rating scale, the test topics, the examinees (undergraduates versus graduates), scoring methods used in the post-rating session, and the five new assessment criteria (See Appendix CC).

First, one rater, R2, mentioned that the range of the scale for ESL 115 was wider because it includes the exempt level for graduates, which might affect rating performance. Second, three raters pointed out that the topic of cloning seemed difficult to write an essay about, and that the content should be updated. R4 pointed out that the topic of globalization should be also updated, and more details and sufficient information should be provided, such as figures and statistics tables. Regarding the examinees, raters mentioned that the level of undergraduate writing seemed quite different from graduate writing; the graduate essays seemed more sophisticated and well-organized, with a higher level of vocabulary. This might affect rating performance, because the same rating scale and descriptors are used for both graduates and undergraduates. Next, raters indicated that both scoring methods were useful. The analytic method was useful to see examinees' strengthens and weaknesses but the holistic method was helpful to decide the final decision. Finally, raters made comments on the assessment criteria. All raters agreed on the importance of the focus criterion, indicating that focus was useful to see whether the essay stayed on topic. Raters showed differing opinions for the criterion of support/elaboration, saying that it was somewhat helpful but not necessarily effective for sorting out plagiarism, because many essays used the examinees' experiences and the lecture session as sources. These findings show that the raters' written reports provided

useful feedback for the modification of the EPT rating system and for the preparation of the next workshop.

Chapter Five

Discussion and Conclusion

Development of the EPT Rater Training Workshop

Rating or grading an essay is a responsibility of language teachers in the language program. For this reason, supervisors of the language program should provide guidelines about the assessment procedures in order to acquire high rater reliability (Fulcher and Davidson, 2007; Lynch, 1996; Phillips, 1997; Rothwell & Kazanas, 2004; Waagen, 2006). Fulcher and Davidson (2007) suggested some ways to improve rater reliability, such as intensive training and sample practice. The focus training program for this study was developed to achieve these goals based on the analysis of several rater training programs, and it seems to have a lot of merits for improving inter- and intra-rater reliability, rating accuracy, and rating validity in terms of authenticity, efficiency, interactiveness, and basis in theory.

First, with respect to authenticity, the training program was designed with context-sensitivity and in collaboration with a practitioner in the field in order to resolve practical concerns. Rating materials were updated and developed to reduce the gap between what was actually happening and what should have been happening. New scoring methods were adopted and authentic materials, such as prototype essay samples and the current version of the reading materials were used in the practice session. A clear EPT rating policy was suggested through the iterative feedback system.

Second, in terms of efficiency, this system is convenient in that raters can easily access the training materials via the internet. It is easy to update and revise new content,

the overall rating system, and the scale descriptors over time based on results of discussion and changes in rating policy.

Third, in terms of interactiveness, this workshop program was designed around interactive feedback. The language teachers asked for help from the trainer when they confronted challenges. The language teachers could get feedback or information about their achievements through face-to-face discussion.

Finally, this workshop program was based solidly in theory because their development was based on program evaluation, language testing, and training theory. The structure, content, and activities of the program were theory driven to promote an accurate understanding of the rating task and shared practical experience with peers, in order to reduce systematic errors in rating. This indicates that the proposed workshop program reflects not only the practical concerns of the practitioners but also theory.

Summary of the Findings

Answers to the four research questions can be summarized as follows:

Findings of question 1. The findings of the pre-workshop survey show that the EPT raters positively evaluated the EPT rating system and the prior training program. Some rater concerns were found based on the closed questions of the pre-workshop survey. First, raters pointed out that the accuracy of the holistic scale descriptors should be improved. This can be considered a source of difficulty in essay rating, since the higher level descriptors may have been less accurate. With respect to test topics, the findings show that individual raters might have different perceptions of the difficulty

level of the topics. It turned out that teaching experience may affect the final decision during the consensus process.

Second, raters wanted to improve the quality of the rating in the training program. Most raters wanted further training on rating skills such as internalization of prototype samples and tips for rating, rather than acquiring knowledge about the rating system. It was suggested that the analytic scoring method be employed for more precise decisions during training. It is interesting that the EPT trainer also suggested this. In addition, the trainer agreed that the focus of training should be modified so that rater reliability and rating accuracy were considered as well as rater agreement. Last, raters wanted to have more feedback from the trainer.

Third, raters had concerns about the rating materials, stating that the training program should be re-organized, the training materials should be updated, and sufficient rating practice should be provided. A new rating system to effectively rate essays on the borderline between two proficiency levels should be designed.

Finally, the EPT trainer echoed the concerns of the raters about the organization of the workshop program, calling for an update to the workshop content and delivery method. The findings from the document analysis confirmed that the trainer, raters, and EPT Gas have limited access to rating materials depending on their responsibilities related to the EPT rating. It was suggested that relevant training materials be collected and synthesized so that the stakeholders can share the materials. It was also suggested that the workshop program allocate more time to the prototype sample practice and feedback. The findings of needs analysis (survey and document analysis) were significant, because they became the interim outcomes (or new input). These findings also yielded

implications for the next stage: workshop preparation. Effort was made to design the new workshop program to address the practical issues/concerns of the raters.

Findings of question 2. Based on the results of the needs analysis, the strengths and weaknesses of the EPT training program were clearly identified, and the ways in which the workshop should be standardized were determined. On the basis of the interim outcomes of the needs analysis, the training content was modified in collaboration with the EPT trainer. The researcher and trainer prepared rating principles based on the prototype samples, focusing on the overall organization of the training program, and a clearer understanding of the rating materials, including the development of analytic scale descriptors and the selection of new prototype essays.

The researcher shared the findings of the pre-workshop survey and the document analysis. To explore the constructs of the EPT rater training program, each of the skills and methods evaluated were identified to enhance the effectiveness of the program. Also, the trainer provided several practical issues with respect to programming the new workshop. The trainer modified a draft summary of the rationale for replacing the training content and methods, including both holistic and analytic scoring methods and prototype essays. The content of the revised training program was modified in consideration of practical concerns.

Findings of questions 3. The proposed workshop was successfully conducted. The researcher observed the workshop session, and the workshop was audio-recorded. All materials and products of the activities were collected and analyzed as evidence. The findings indicate that further training or practice might be necessary to internalize the prototype samples across proficiency levels. In addition, it seemed that raters were highly

motivated to share their knowledge and rating experiences to resolve their concerns and challenges.

After the workshop, the strengths and weaknesses of the proposed training program was assessed through empirical evidence. The findings suggest that raters were overall satisfied with the goals, schedule, content, delivery method, feedback, and activities. The confidence level of the new raters gradually increased after the workshop in comparison with that of the experienced raters. These results show that the workshop program was effective for enhancing rater reliability and rating validity.

Conversely, the workshop duration might be improved, although opinions were controversial. Some raters felt that the three-hour workshop was too short to cover the program, but some pointed out that it was too long, compared to previous workshops. It was suggested that the procedures for Activities 1, 2, and 3 be modified for the next workshop, because time for each activity had not been evenly assigned. Finally, it was suggested that the analytic scoring guidelines be revised.

Findings of question 4.

Effects of the systematic training program. The EPT training was innovatively reorganized and upgraded. The training program was designed to take advantage of four different types of training program (Woehr & Huffcutt, 1994), and focused on enhancing rater consistency and accuracy, as well as rating validity by reducing systematic errors (McNamara, 1996; Furneaux & Rignall, 2007; Shaw, 2002). The trainer and the EPT G.A. integrated and updated the training materials to fit the current rating context. Prototype essays were re-selected in consideration of the focus of activity for each phase of the proposed training program.

The training program was designed so that raters could perform evidence-based judgments and solve their rating problems. The workshop lecture provided substantial information, and raters learned how to accurately observe the essay for the appropriate information to match evidence from the rating scale descriptors, and finally how to make a final decision.

It was significant that the summary of the findings of the workshop activity showed evidence that raters successfully understood the training focus and followed the directions. Moreover, a contextualized feedback system was employed via group discussion. Feedback from peers and the trainer was instantly given; however, the consensus process was slightly different depending on the peers. Some groups conducted the consensus process successfully, but some did not. It was suggested that the lecture session provide information about the evidence-based decision process.

Effects of test topic/prompts. The findings of this study confirmed the findings of Hamp-Lyons & Mathias (1994), and Shaw & Weir (2007) in which raters have different perceptions of test topic difficulty, although it was not statistically significant in this study. The findings supported Weigle's study (1994a) in the sense that rater performance was different depending on the test topic.

On the basis of the investigation of the rating split rate for 2009 and Spring of 2010, the test topic seems to have some influence on rating performance, particularly on inter-rater reliability (agreement). The findings show that a larger rating split was found for the topic of cloning across the two years compared to the other topics, although the rating split rate was lower in Spring of 2010. However, based on the FACETS analysis of reliability, no significant test topic effect was detected, and rater severity level was

similar across test topics, with only three significant cases in which rater and test topic interaction was detected.

Test topic did not seem to affect rating performance in terms of intra-rater reliability and severity level; however, test topic might affect rating accuracy or rating agreement in some way. In addition, it is likely that, although some raters discriminated among topic difficulty (Shaw & Weir, 2007), they did not rate more leniently for the test topics they perceived as difficult. It can be said that the training contributed to increasing inter-rater reliability, with a decreased rating split after the workshop.

Effects of scoring method and assessment criteria. It was significant that scoring method may be a strong candidate affecting rater severity, rating accuracy, agreement, and interaction effects. Both holistic and analytic scoring methods were employed under the training situation, and the five assessment components were presented. It was pointed out that new analytic guidelines may guide raters to easily reach a final decision (Shi, 2001). It was interesting that no rater bias for assessment criteria was detected, but the analytic score method may lead raters to rater more harshly (Cumming, 1990).

In terms of comparison among the three raters, the overall severity level of the workshop group was similar to the control group based on the means of the holistic score. The workshop group raters scored relatively more leniently than the control group for the analytic scoring. The experienced raters scored relatively more severely than the new raters for both holistic and analytic scoring, and the standard deviation indicated that experienced raters used a wider rating scale, avoiding a central tendency. Rating accuracy might be variable depending on the scoring method across test topics. For the holistic rating, the rating accuracy for the topic of globalization was the highest, and accuracy

was low for the topic of animal testing. For the analytic scoring, rating accuracy for the topic of animal testing was the highest, and rating accuracy for the topic of globalization was the lowest.

Regarding proficiency level, when the analytic scoring method was used, raters scored more accurately at the lower levels, such as ESL 113 or ESL 500, and for undergraduate students, rather than graduates. Conversely, it seemed that the holistic scoring method may be more effective to score the higher proficiency levels and graduate student essays, although there was variability among raters. This result showed counterevidence for the results of Cumming's study (1990).

Finally, variance analysis using GENOVA supported these results, because a large amount of variance was estimated with P x A x R, and scoring method might be considered to affect rating performance. FACETS analysis showed more misfit essays with the analytic scoring method. The difference in the severity level was small for analytic scoring, indicating that the analytic method may reduce severity among raters.

Effects of examinee proficiency level. The findings of this study partly support the findings of Kondo-Brown (2002) and Schaefer (2008), in which raters may have different perceptions of examinee proficiency level. Based on the raters' comments, raters perceived that graduate level essays showed a higher quality of academic writing than those of undergraduates, and it seemed that raters might have different standards for rating the two groups. However, graduate versus undergraduate status did not affect the rating process, and evidence of contrast effect could not found. It is likely that raters need more training to rate borderline essays between ESL 114 and ESL115 and between ESL 501 and ESL 500. Finally, test topics were scored consistently with a similar severity

level, and the five assessment criteria were also consistently scored, indicating that they might be considered appropriate criteria for assessing the EPT essays.

Rating accuracy and rater reliability with respect to rater background. Accuracy was estimated by matching the raters' scores with the trainer's score from the workshop activities, and with the original essay scores in the post-rating session. Based on the results of the post-rating session, accuracy was somewhat low and rater accuracy should be improved. Based on the results of the post-rating session, rater reliability was explored with a quantitative approach by identifying systematic error. Using GENOVA, the variance of rater aspect was estimated, and the findings indicate that the rater effect was minor regardless of scoring method. It is interesting to note that the large amount of interaction among facets accounted for the majority of the errors. Based on FACETS analysis, rater consistency was calculated, and two raters (R1 and R4), who was in the control group, showed an overfit or misfit rating pattern, although their severity was diverse, from -3.16 to -2.09 logit, based on the holistic scoring. However, in the analytic scoring, four raters (R3, R4, R5 and R6) showed an overfit or misfit rating pattern.

According to the analysis of Cohen's kappa, all six raters showed low agreement which supports the findings of the simple percentage of agreement. R1, R2, and R5 showed a fair level of agreement with the original holistic score, while the remaining three raters showed low agreement with the original holistic score. FACETS analysis supported this result that the inter-rater reliability among the six raters was quite low, showing 52.5% on average when the holistic scoring method was used, or 20.6% agreement when analytic scoring was used. In group comparisons, the workshop group scored more accurately, and the new raters rated more accurately than experienced raters.

This indicates that teaching/rating experience did not critically influence rating performance. Finally, neither halo effects on the basis of the results of the correlation analysis, nor contrasts or similarity effects based on the FACETS bias analysis were detected; however, further study should be conducted regarding halo effects after revising the analytic scoring guidelines.

Expert rating and rating validity. This study tried to define the expert rating pattern and strategy in more detail. Based on the quantitative findings of the post-rating session, raters showed a good rating performance in terms of intra-rater reliability, but rating accuracy and agreement were quite low. The nature of expert rating can be discussed. Normally, intra-rater reliability is an index used to judge rating quality, but we need to think about whether this standard can be applied to measure training effectiveness.

The results of the qualitative analysis of the raters' consensus processes confirmed that expert raters conducted more structured reading. Expert raters tried to persuade their peers using evidence from the rating scale and essays rather than intuition. The findings supported the studies of Feldman (1981), Pulakos (1984), and Eckes (2008), in which the information processing model was standardized and contributed to the decision-making process. This serves as evidence that expert raters accurately observed, selected, and categorized information from the essays and scale descriptors. The final judgment was reasonable, formed by considering the EPT writing courses and individual experience. Expert raters tended to use not only top-down but also bottom-up approaches, scoring holistically to start, and then finding local evidence with the analytic approach (Brown, 1995; Weigle, 1994a). This indicates that analytic scoring may be useful for selecting evidence to support a global impression.

However, when the analytic score conflicted with the holistic score, the expert raters tended to consider the holistic score as more important. The results consistently supported the findings of Bernardin and Walter (1977) that raters showed personal preference on the type of rating scale, with expert raters tending to prefer simple scale descriptors.

Conversely, unskillful raters consistently showed the following characteristics, which can be explained in terms of the standardization of the internal rating process. First, they did not fully understand the essay or the rating scale descriptors. Second, they tended miss meaningful facts or evidence. Although they successfully found evidence from the essay, they had problems, for example, in linking the evidence with the scale descriptors. Unskillful raters tended to use a bottom-up approach in which they recalled and retrieved specific facts, so they may have had difficulty weighting/valuing at the final judgment stage. They might have made the final decision based on one aspect of the essay, or failed to make the correct decision.

It is interesting that rating experience did not seem critical for having an expert rating pattern, but it is likely that the internalization of a series of processes, such as correct observation, information selection, categorization and judgment, may be associated with expert rating in this study. Moreover, more teaching/rating experience sometimes may negatively influenced rating because rater bias was not improved with the one-day workshop. This indicates that teaching/rating experience does not always guarantee expert rating, and this result supports previous studies (Cumming, 1990; DeRemer, 1998; Ersosy, 2004; Homburg, 1984).

Regarding decision making strategy, raters used the rating strategies shown in Figure 5. Strategy at the observation and judgment stages was clearly distinguishable, but raters seemed to use the rating strategies of selection and categorization simultaneously. They adapted different strategies depending on the rating context. Rating validity was evaluated by estimating inter-and intra-rater reliability, rating accuracy, and analysis of expert rating patterns. The study suggested that group discussion, feedback, and practice of evidence based judgment were good methods for enhancing rating validity and reliability. Through the group dynamic, individual raters developed an awareness of their own rating patterns by comparing them with their peers. Severity level might be effectively adjusted based on group discussion. Whenever unskillful raters confronted a problem during the cognitive information process, they had the opportunity to learn how to resolve the problem with contextualized peer feedback.

Implications of the Study

An effort was made to enhance the quality of the EPT training program and rater performance. The new training program focused on addressing stakeholder concerns from a practical perspective, informed by theoretical issues. The findings of this study will contribute to the design of a better training program in the future. The content of the lecture, the scoring methods, and the group activities should be updated and effectively implemented in the future. In addition, an expert rating pattern was partly identified, which will yield implications for future workshops. Expert rating skills will be described via lecture and used by raters who have rating difficulties. It seemed that raters who participated in this study were highly motivated by receiving contextualized feedback and

encouragement from the trainer. Also, they accelerated their professional development with high self-awareness. Specifically, the findings of this study are important to understanding the rater training program and resolving its practical concerns. Additionally, this study is expected to yield valuable insights for designing a rater training program.

Limitations of the Study

There are several limitations of this study. The first limitation was time constraints. Because of a lack of time, more detailed explanations and examples for each level, especially analytic scoring guidelines, could not be provided. The lecture should have sufficiently covered how to conduct evidence-based decision making, and raters should have had more prototype essays to rate. In addition, the analytic guidelines and scoring system should be revised through further study. Second, although an analytic scoring was employed to enhance rating validity and rating reliability, it had not been perfected. It needs to be modified for the next workshop program.

Third, content analysis of the essays was not conducted. Based on the bias analysis of FACETS, essays which showed interaction effects with raters were not analyzed in this study. Content analysis of the essays is required for exploring a source of rater variability in terms of what essay features affect rating performance.

Fourth, originally, both pre- and post-rating sessions was proposed, but due to the difficulty of recruiting participants, the research design was modified to include only the post-rating session. It would have been helpful to compare the quality of rating before and after the workshop, and the evidence would contribute to estimating the effectiveness

of the training. Thus, in this study, training effectiveness and data interpretation were somewhat unclear because rater improvement was not clearly identified. For future study, it is suggested that better measures be adopted and implemented.

Finally, sampling was one of the limitations of this study. Different participants from the same population were recruited at different research phases. Particularly, for the post-rating session, five of the six raters were concurrently serving as ESL teachers, and they had been involved in additional rating sessions for classroom diagnostic testing. It is likely that practice effects from this additional rating may have reduced training effectiveness, because the one rater in the control group who was not a teacher showed a clearly misfit rating pattern. In addition, a small number of participants was recruited for the pre-workshop survey (n=8), workshop session (n=15), and post-rating session (n=6), which makes it difficult to generalize the results across contexts.

Suggestions for the EPT Workshop

The current EPT rater training program was evaluated and modified in collaboration with the EPT trainer. There are several suggestions for future workshops. First, based on the analysis of the audio-recorded training workshop, the lecture session simply delivered the content of the existing materials, such as test topics and rating procedures. However, as raters have already pointed out in the post-workshop survey, more detailed explanations and appropriate examples of proficiency levels and assessment criteria should be presented. Detailed explanations with prototype essays would be helpful for raters to engage in evidence-based decision making and the group consensus process. Additionally, the training should focus on internalization of prototype

essays and more structured reading and rating steps suggested in this study, rather than relying on teaching/rating experience. Regarding training focus, rating accuracy should be given more consideration over rater agreement and consistency.

Second, the rating scale descriptors, especially the analytic scoring guidelines, should be revised. Its numerical scoring system should be adjusted based on empirical data, and the descriptors should also be reviewed by raters. Another suggestion is that, since misfit essays based on FACETS analysis were found, misfit essays should be discussed with raters during the workshop session. It is expected that such discussion may provide ideas as to how to handle difficult essays. Finally, if possible, more rating practice should be conducted. Based on the analysis of the audio-recordings, time for the three activities was not evenly assigned. For instance, Activity 1 took a lot of time. The time management for each workshop step should be carefully considered (see Figure 20).

Major concerns	Attempts	Improved area for the next workshop
Organization of the workshop program	• Re-organized the program • Clarify the training focus • Integrated training material via web-tool	• Lecture session should be strengthened by providing more explanation. • Agreement and rating accuracy should be improved.
More practice and lack of teaching experience	• Three activities were implemented for different purposes. • Group activity was implemented to cover lack of teaching experience.	• Timelines for each activity should be improved and have more practice.
Borderline essay	• Analytic scoring method was employed	• Rating scale and its descriptors should be modified because raters were not accustomed to the analytic system. And it is a source that rater tended to be severe.
Consistency, accuracy, and agreement	• Various activities were conducted to improve these aspects.	• Rating accuracy should be improved.

Figure 20. Suggestions for the future training program.

Suggestions for Future Research

There are three suggestions for the future research. First, when research related to rater training effectiveness is conducted, a better research design is required to estimate training effectiveness. During this study, raters had many inputs in the workshop session, so sometimes it was difficult to estimate and interpret the workshop effects. Second, scoring methods affect rater performance, whether positively or negatively. From a practical perspective, it should be identified through future research which scoring method would be more useful for to train raters and help them internalize rating skills. Finally, regarding rating accuracy and rater reliability, the nature of expert rating should be also investigated. The relationship between rating accuracy, agreement, and consistency should be explored. Also, it should be studied further which of the three is most important in training, and how to improve the rating performance in consideration of these aspects.

Closing Remarks

The rationale behind this study was that low rater reliability would be a strong candidate to threaten test usefulness in terms of test validity and practicality as well as reliability. Standardization of the training program was conducted and evaluated. However, it is too early to clearly answer whether the new EPT training program improved rating quality in terms of rater reliability and rating validity or not. The training program will continue to be revised and validated for improvement. When I started this research, I believed that a more systematic training program would contribute to enhancing the quality of rating, and if an expert rating pattern could be clearly identified,

I think this also might yield great insight for rater training. I still hold the opinion that a more systematic training program is helpful to improving the quality of rating, but some unsolved issues still exist. I will conduct follow-up research based on the findings of this study to try to resolve practical concerns on the basis of theory.

References

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale, *Biometrics, 44*(2), 539-548.

Agresti, A. (1996). *An introduction to categorical data analysis*, New York, NY: A Wiley-interscience Publication.

Agresti, A., & Winner, L. (1997). Evaluating agreement and disagreement among movie reviewers, *Chance, 10*(2), 10-14.

Alderson, J. C. (1991). Bands and scores. *Review of English Language Testing Teaching 1*(1). 71-86.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, California: Brooks/Cole Publishing Company.

American Council on the Teaching of Foreign Languages (ACTFL). Retrieved from http://www.actfl.org,.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing,* Upper Saddle River, NJ: Prentice –Hall.

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work. *Annual Review, 49,* 141-168.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology, 72*(4), 567-572.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10,* 149-164.

Bachman, L. F. (1990). *Fundamental considerations in language testing,* Oxford, UK: Oxford University Press.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*, Cambridge, UK: Cambridge University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12,* 238-257.

Bachman, L. F, & Palmer, A. (1996). *Language Testing in Practice*, Oxford UK: Oxford University Press.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal, 70(5),* 380-390.

Berlitz Proficiency Interview. (2006). Unpublished BPI Rater Training Materials. Berlitz Inc.

Bernardin, H. J., & Buckley, M. R. (1981). A consideration of strategies in rater training, *Academy of Management Review, 6,* 205- 212.

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62*(1), 64-69.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 65*(1), 60-66.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60*(5), 556-560.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behaviour and Human Performance, 20,* 238-252.

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63*(2), 135-144.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64*(4), 410-421.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12,* 1-15.

Caracelli, V. J., & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 15*(2), 195-207.

Cardy, R. L., & Kehoe, J. F. (1984). Rater selective attention ability and appraisal effectiveness: The effect of a cognitive style on the accuracy of differentiation among ratees. *Journal of Applied Psychology, 69,* 589-594.

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the Irrelevant: Anchors in Judgments of Belief and Value, In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment.* Chapter 6, 120-138. Cambridge, UK: Cambridge University Press.

Choi, Y. H. (2000). Effects of writing test tasks on learner performance and rating. *English Teaching, 55*(3), 217-245.

Choi, Y. H. (2002). FACETS Analysis of effects of rater training on secondary school English teachers'scoring English writing. *Journal of the Applied Linguistics Association of Korea. 18*(1), 257-292.

Cooper, W.H. (1981a). Conceptual similarity as a source of illusory halo in job performance ratings. *Journal of Applied Psychology, 66*(3), 302-307.

Cooper, W.H. (1981b). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218-244.

Cooper, W.H. (1983). Internal homogeneity, descriptiveness and halo: Resurrecting some answers and questions about the structure of job performance rating categories. *Personnel Psychology, 36,* 480-502.

Crow, W.J. (1957). The effect of training upon accuracy and variability in interpersonal perception. *Journal of Abnormal and Social Psychology, 55,* 355-359.

Charney, D. (1984). The validity of using scoring to evaluate writing: A critical overview. *Research on the Teaching of English, 18*(1), 65-81.

Cumming, A. (1990). Expertise in evaluating second language compositions, *Language Testing, 7*(1), 31-51.

Cumming, A., Kantor, R., & Power, D.E. (2002). Decision making while rating ESL/EFL writing tasks: a descriptive framework, *The Modern Language Journal, 86*(1), 67-96.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT and London, UK: Yale University Press.

DeRemer, M. L. (1998). Writing assessment: raters' elaboration of the rating task. *Assessing writing, 5*(1), 7-29.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Elder, C., Barkhuizen, G., Knoch, U., & Randow, J. (2008). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37-64.

Elder, C., Knoch, U., Barkhuizen, G., & Randow, J. (2008). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*(3), 175-196.

Engelhard, G. J. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Erdosy, M. U. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. TOEFL Research report, RR-03-17

ESL Placement Tests (EPT) at University of Illinois. Retrieved from http://www/linguistic.illinois.edu/students/placement/

Freedman, S.W., & Calfee, R.C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, S. Walmsley (Eds.) *Research on writing: Principles and Methods* (pp. 75-98). New York, NY: Longman.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*(2), 127-148.

Fulcher, G. (1987). Tests of Oral performance: the need for data-based criteria. *English Language Teaching Journal, 41*(4), 287-291.

Fulcher, G., & Davidson, F. (2007). Language Testing and Assessment. New York, NY: Routledge

Furneaux, C. & Rignall, M. (2007). The effect of standardization-training on rater Judgments for the IELTS writing module. In M. Milanovic & C. Weir (Series Ed.) and L. Taylor and P. Falvey (Vol Ed.), *Studies in Language Testing 19, IELTS Collected papers, research in speaking and writing assessment* (pp. 422-444). Cambridge UK: Cambridge University Press.

Gilbert D. T. (2002). Inferential Correction, In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment.* Chapter 9, 167-184. Cambridge, UK: Cambridge University Press.

Goodenough, F. L. (1950). Mental Testing. New York, NY: Rinehart Ncompany Inc.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework evaluation for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis. 11*(3), 255-274.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining Expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3*(1), 49-68.

Haswell, R. H. (1998). Rubrics, prototypes, and exemplars: categorization theory and systems of writing placement. *Assessing Writing, 5*(2), 231-268.

Hauenstein, N. M. A. & Foti, R. J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. *Personnel Psychology, 42*, 359-378.

Henning, G. (1987). A guide to language testing; development, evaluation, research. Boston, MA: Heine & Heine Publishers.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: can it be validated objectively? *TESOL Quarterly, 18*(1), 87-107.

Jang, S. Y. (2006). The Development of an Effective Rater Training Model for ESL Teachers. Unpublished Early Research, University of Illinois, Urbana-Champaign.

Jang, S. Y. (2007). Unpublished review paper answering a specific qualifying exam, University of Illinois, Urbana-Champaign.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S.R., & Fisher, S.P. (2005). Resolving score differences in the rating of writing samples: does discussion improve the accuracy of scores*? Language Assessment Quarterly, 2*(2), 117-146.

Kahneman, D. & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment, In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge, UK: Cambridge University Press

Keysar, B., & Barr, D. J. (2002). Self-Anchoring in Conversation: Why Language Users Do Not What They "Should", In T. Gilovich, D. Griffin, & D.Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment,* Cambridge, UK: Cambridge University Press

Kiely. R., & Rea-Dickins, P. (2005). *Program evaluation in language education*, New York, NY: Palgrave Macmillan.

Kim, J. Y. (2008). Unpublished Doctoral Dissertation. University of Illinois, Urbana-Champaign, Development and validation of an ESL diagnostic reading-to-write test: An effect-driven approach.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing 19*(1), 3-31.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize ratings errors in the observation of behavior. *Journal of Applied Psychology, 60*(5), 550-555.

Lee, C. (1985). Increasing performance appraisal effectiveness: matching task types, appraisal process, and rater training. *Academy of Management Review, 10*(2), 322- 331.

Levine, J., & Buttler, J. (1952). Lucture versus group discussion in changing behavior. *Journal of Applied Psychology, 36*, 29-33.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: Mesa Press, University of Chicago.

Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency model? A quasi-experiment. *Journal of Applied Psychology, 92*, 812-819.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70*(1), 66-71.

Lord, F. M & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley Publishing Company.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19*(3), 246-276.

Lumley, T. (2005). Assessing Second Language Writing: The Rater's Perspective, Frankfurt, Germany: Peter Lang GmbH.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Lunz, M. E, Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity in examination scores. *Applied Measurement in Education, 3*(4), 331-345.

Lynch, B.K. (1996). Language Program Evaluation. Cambridge, UK: Cambridge University Press.

Lynch, B.K. (2003). Language Assessment and program evaluation. Edinburgh, Scotland: Edinburgh University Press.

Lynch, B.K., & McNamara, T.F. (1998). Using G- theory and many-facet Rasch measurement assessment in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158-180.

Mathison, S. (1992). An evaluation model for inservice teacher education, *Evaluation and Program Planning, 15*, 255-261.

McIntyre, R. M., Smith, D., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*, 147-156.

Multi-media Assisted Test of English (MATE). (2006). Unpublished MATE Rater training handbook. Sookmyung Women's University Press, Seoul.

McNamara, T. F. (1996). *Measuring Second Language Performance.* London, UK: Longman.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic (Series Ed.) & M. Milanovic and N.Saville (Vol Ed.), *Studies in Language Testing 3, Performance Testing, Cognition and assessment, selected papers from the 15th language testing research colloquium, Cambridge and Arnhem* (pp.92-114). Cambridge, UK: Cambridge University Press.

Mislevy, R. J. (2004). Can there be reliability without "reliability?", *Journal of Educational and Behavioral Statistics, 29*(2), 241-244.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Research, 23*(2), 5-12.

Moss, P. A. (2004). The meaning and consequences of "Reliability", *Journal of Educational and Behavioral Statistics, 29*(2), 245-249.

Murphy, K. R., & Balzer, W. K. (1989). Rating errors and rating accuracy. *Journal of Applied Psychology, 74*(4), 619-624.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In M. Milanovic & C. Weir (Series Ed.) and L. Taylor and P. Falvey (Vol Ed.), *Studies in Language Testing 19, IELTS Collected papers, research in speaking and writing assessment* (pp. 422-444). Cambridge, UK: Cambridge University Press.

Parkes, J. (2007). Reliability as argument, *Educational measurement: issues and practice, 26*(4), 2-10.

Payne, J. W., Bettman, J. R., & Johnson, E.J. (1993). Contingencies indecision making, *The Adaptive Decision Maker,* (Chapter 2, 20-69), Cambridge, UK: Cambridge University Press.

Phillips, J. (1997). Handbook of Training Evaluation and Measurement Methods. Houston, TX: Gulf Publishing Company.

Pulakos, E.D. (1984). A comparison of rater training programs: error training and accuracy training. *Journal of Applied Psychology, 69*(4), 581-588.

Reid, J. (1993). Historical Perspectives on Writing and Reading in the ESL Classroom, In J. Carson & I. Leki (Eds.), *Reading in the Composition Classroom; Second Language Perspectives,* (Chapter 2, 33-60). Boston, MA: Heinle & Heinle Publishers.

Roch, S. G., & O'Sullivan, B.J. (2003). Frame of reference rater training issues: recall, time and behavior observation training. *International Journal of Training and Development, 7*(2), 93-107.

Rothwell, W.J., & Kazanas, H. C. (2004). Mastering the Instructional Design Process. San Francisco, CA: Pfeiffer.

Saal, F .E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings; Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2),413-428.

Salaberry, R. (2000). Revising the revised format of the ACTFL oral proficiency interview. *Language Testing, 17(3),* 289-310.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Shaw, S.D. (2002). The effect of training and standardization on rater judgment and inter-rater reliability, *Research Notes 8,* 13-17.

Shaw, S.D., & Weir, C. J. (2007). Examining Writing. In M. Milanovic & C.Weir (Series Ed.) and S. Shaw.& C. Weir (Vol Ed.),Studies in Language Testing 26, Research and practice in assessing second language writing. Cambridge, UK: Cambridge University Press.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing, *Language Testing, 18*(3), 303-325.

Shin, D. I. (2001). Exploring rating patterns with Rasch measurement techniques: implications for training. *Foreign Languages Education, 8*(1), 249-272.

Shin, D.I., & Jang, S.Y. (2002). Understanding rating error sources on halo effect. *Foreign language Education, 9*(4), 215-232

Shohamy, E. (1990). Language testing priorities: a different perspective. *Foreign Language Annuals, 23,* 385-394.

Shohamy, E., Gordon, C., & Kraemer, R.(1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76* (1), 27-33.

Slusher, M.P., & Anderson, C.A. (1987). When reality monitoring fails: The role of imagination in stereotype maintenance. *Journal of Personality and Social Psychology, 52*, 653-662.

Smith, D.E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review,11*(1),22-40.

Spool, M. D. (1978). Training programs for observers of behavior: A review. *Personnel Psychology, 31*, 853-888.

Steward, S. (1999). Unpublished Master Thesis, University of Illinois at Urbana – Champaign, An investigation of writing features which have predictive evidence of validity for EPT placement levels.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns, *Journal of Applied Psychology, 73*, 497-506.

Sulsky, L. M., & Day, D.V. (1992). Frame-of-reference training and cognitive categorization:an Empirical investigation of rater memory issues. *Journal of Applied Psychology, 77*(4),501-510.

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology, 65*, 351-354.

Tversky, A., & Kahneman, D. (2002). The conjunction Fallacy in Probability Judgment, In T. Gilovich, D. Griffin & D. Kahneman (Eds), *Heuristics and Biases: The Psychology of Intuitive Judgment*, Chapter 1, Cambridge, UK: Cambridge University Press.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal*, *49*(1), 3-12.

Varghan, C. (1993). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Eds.), *Assessing Second Language Writing*, Chapter 6, 111-125. Norwood, NJ: Ablex Publishing Corporation.

Waagen, A.K. (2006). Infoline Guide to Training Evaluation. Alexandria, VA: ASTD Press.

Warmke, D. L., & Billings, R.S. (1979). Comparison of training methods for improving the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

Weigle, S. C, (1994a). Unpublished Doctoral Dissertation. University of California, Los Angles, Effects of training on raters of English as a second language composition: Quantitative and qualitative approaches.

Weigle, S. C. (1994b). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weiss, C. H. (1998). Evaluation. Upper Saddle River, NJ: Prentice –Hall.

Wexley, K. N., & Latham, G. P., (2002). Developing and training human resources in organization. Upper Saddle River, NJ: Prentice Hall.

Wiggleworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction, *Language Testing 10*(3), 305- 335.

Wildman, B. G., Erickson, M.T., & Kent, R. N. (1975). The effect of two training procedures on observer agreement and variability of behavior ratings. *Child Development, 46,* 520-524.

Wilson, T. D., Centerbar, D. B., & Brekke, N. (2002). Mental Contamination and the Debiasing Problem, In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. Chapter 10,185-200. Cambridge, UK: Cambridge University Press.

Woehr, D.J., & Feldman, J. (1993). Processing objective and question order effects on the causal relation between memory and judgment in performance appraisal: The tip of the iceberg. *Journal of applied Psychology,78*(2), 232-241.

Woehr, D.J., & Huffcutt, A. I,(1994), Rater training for performance appraisal: A quantitative review, *Journal of Occupational and Organizational Psychology, 67,* 189-205.

Appendix A

Evaluation of Documentation for the Current Rater Training Program: Four Tests

Stage	Familization	Norming				Additional Information	
Content/activity	Lecture	Practice	Sample speech/ rating process	Evaluation	Policy for the rating	Supplementary information	Technical information
ACTFL OPI	√	√	Insufficient	×	√	×	×
MATE	√	√	√	Insufficient	√	√	√
BPI	√	√	Insufficient	×	Insufficient	×	×
EPT	Insufficient	√	√	×	√	×	×

Appendix B

Strengths and Weaknesses of Several Training Programs and Suggestions for Improvement for Each

Tests	Strengths	Weaknesses	Suggestions for improvement
ACTFL OPI	▪ Well-structured training program ▪Principles and strategies for interviewing and decision making ▪Rating policy for interview and scoring	▪ Lecture style ▪More activities for simulation interview session ▪Lack of the study on the quality of speech samples, and an interviewer behaviors ▪Insufficient supplementary information ▪ Evaluation and feedback	▪ Intensify the norming ▪Provide more authentic activities reflecting interview context ▪Provide more detailed information on the descriptors ▪Provide supplementary information on the rating policy and raters' certification process ▪ Need two-way feedback on mutual evaluation
MATE	▪ Well-structured training program ▪ Visual-aids used ▪Principles and strategies for rating provided ▪Supplementary information	▪ Lecture style ▪ Insufficient activities for norming ▪Lack of the study on the quality of speech samples, and an interviewer behaviors	▪ Intensify the norming ▪ Provide more authentic activities ▪Provide more detailed information on the descriptors ▪ Need two-way feedback on mutual evaluation
BPI	▪ Well-structured training program ▪Explanation on general/ specific information ▪ Test specification based	▪ Lecture style ▪ Insufficient activities using actual samples ▪More activities for simulation interview session ▪Lack of the study on the quality of speech samples, and an interviewer behaviors ▪Unclear rating policy ▪Insufficient supplementary information ▪ Evaluation and feedback	▪Questioning and activity reflecting interview context ▪ Intensify the norming ▪Provide more detailed information on the descriptors ▪Provide supplementary information on the rating policy and raters' certification process ▪ Need two-way feedback on mutual evaluation
EPT	▪Practice focused rather than lecture ▪ Rater centered materials ▪Deep understanding the essay samples of test takers	▪Short training session and irregular workshop ▪Dependence on TA experience ▪Lack of general principles and strategy ▪Insufficient supplementary information ▪ Evaluation and feedback	▪ More systematic and regular training program ▪ Provide general principles and strategies ▪Provide supplementary information on the rating policy and raters' certification process ▪ Need two-way feedback on mutual evaluation

Appendix C

Arguments for the Main Study

Describing the stages of the research phase and evidence presented from each stage of Context Adaptive Model by Lynch (1996;2003)

Phase	*Claims*	*Assumptions*	*Evidence & Sources of Evidence*	*Methods of Analysis*
Step 1 & 2 (The distinction between steps (1) and (2) is clarified in the next table.)	That knowledge of the evaluation context and the specific program to be evaluated will be identified.	i) Training program and stakeholders' concerns should be identified in line with the testing needs or raters' tasks. ii) Setting goals and analysis of audience	1. Needs analysis is conducted to understand the program being evaluated. i) Document analysis ▪ Analysis of the EPT specification ▪ Analysis of the current training program's materials and current EPT writing test ii) Survey analysis of stakeholders' views toward BPI training program.	i) Qualitative/quantitative analysis of needs of stakeholders and the program (questionnaire) ii) Document analysis of current training program (materials, activities, and samples for practices) for all sections.
Step 3	That the construct of the BPI rater training program is identified based on a strong theoretical framework for assessing raters' performance, each of the skills evaluated, and evaluating the effectiveness of the program.	The model of the EPT training program is designed and standardized based on extensive research from three sources: i) Literature review of proficiency based on evaluation theory, training theory, and theory of language testing. ii) Comparisons with other available training programs.	i) Research document on effectiveness of the rater training program to building rationale (program evaluation theory, Test specification theory) ii) Market competition document iii) Interim outcomes from needs analysis	Mapping research/current theories regarding training program design, evaluation and human resource training
Step 4	That the training program can be standardized with	i) Proposed new model for a training program ensures the effectiveness and	i) Process of reverse engineering of existing current rater training program.	i) Qualitative and quantitative analysis of training materials, skill

	regard to the stakeholders' concerns and expectations, and that the strengths and weaknesses of the training program can be assessed through substantive theory.	appropriateness of enhancing the reliability to be assessed. ii) Proposed model provides guiding materials for knowledge, skills, self-awareness and motivation, activities in describing the operational features of the training program. iii) Proposed model provides interim outcomes as well as positive washback effects on raters and the program level.	ii) Conscious effort toward crafting new versions of the standardized training program, and materials (including on-line and on-site materials) and selection of prototype samples and activities (practice, discussion, feedback) produced from different revisions.	based assessment on and rating results
Step 5	That diligence during the training phase itself (for the training method, practices, feedback system) provides a powerful way for raters to understand and enhance their reliability and at the same time provides validity evidence for the quality of rating.	i) Sound practice of rating skills through peer and group feedback as well as expert ratings from a trainer ii) Supporting their problem solving process during rating, and maintaining current skills acquired.	i)Observation of training workshop ii)Results of rating practices iii) Questionnaire provides an argument structure for evaluating the standardized training workshop phase for raters.	i)Qualitative interpretation of field notes ii)Quantitative analysis on rating scores from their practices and survey

Step 6& 7 (The distinction between steps (6) and (7) is clarified in the next table.)	That effectiveness of the training program can be estimated through a process of consensus-based decision.	The proposed more systematic training model is useful for enhancing rater proficiency, and a helpful tool for addressing their concerns.	i) Rater motivation, self awareness, rater reliability/accuracy go up. ii) Rater knowledge about their tasks and their rating skills are enhanced. iii) Raters' judgment process can be monitored via more systematic tool improving the accuracy of rating. iv) More systematic training program supports raters and addresses their concerns.	i)Summary of findings from each phase ii)Evaluation Report

Appendix D

Data Collection Protocol

Activity		EPT Raters: Writing Raters	EPT Trainers	Research Activity (time/place)	Data Collected
1. Needs Analysis	1.1Document analysis			Documents will be collected from the website or provided by the EPT GA.	Documents will be de-identified and included in data analysis.
	1.2 Survey	Raters will be involved in the survey	Rater trainer will take part in the survey.	Survey will take approximately 1 hour and be provided by the researcher along with a self-addressed stamped envelope to return directly to the researcher.	Results of the survey will be anonymous.
	1.3 Writing sample	The writing samples will be obtained from the EPT bank of approved writing samples. Topics chosen will vary providing different rating opportunities.			Writing sample provided will be anonymous and rater will provide rating directly to the researcher by e-mail.
	1.4Pre-rating session: Raters will score and self-audio-tape their scoring strategy.	Writing raters will be asked to score writing samples and will make a retrospective written report on their scoring strategy.		i) Rater will be asked to score a selected essay which may take 5-6 hours. Researcher will provide essays to the rater – to respond at his/her convenience. Score will be collected in person or by e-mail. ii) Raters will do written report. It takes approximately 10-20 minutes in order to provide scoring strategies to researcher.	Scoring results and de-identify audio-recordings of scoring strategies provided by raters.
2.Rater Training workshop	2.1 Preparation		Rater trainer will meet with researcher to revise the workshop.	Researcher will provide a summary of the data analysis to rater –trainer. They will collaborate on developing the training workshop revising curriculum as needed. This step may take several hours over one week- approximately 6 hours in total.	Data collected will be revised workshop materials compared to previous workshop materials.
	2.2 Workshop: This is a required activity for raters but the	Raters will participate in the workshop.	Rater trainers will lead the workshop	i) Workshop will be 6 hours long and audio-taped. ii) A post-workshop survey will	Audio-tapes will be transcribed using codes and surveys will be anonymous.

	research portion will be optional. Only the post workshop survey will require any additional time.			be provided to raters. Survey will take approx 20 minutes and distributed and collected by the researcher.	
3.Post-rating session	3.1 Writing sample 3.2Post-rating session and rater's written report of scoring strategy.	Score the writing assessment as in step 1.		Same activity and data collected as outlined in section one for writing and oral assessment.	Same activity and data collected as outlined in section one for writing and oral assessment.

Appendix E

Participant Consent Forms

Dear EPT Raters:

I would like to invite you to participate in a research project that examines the development and evaluation of a standardized rater training program for the ESL Placement Test (EPT) raters. This research is being carried out by So-young Jang for a doctoral dissertation under supervision of Professor Fred Davidson in the department of Linguistics at University of Illinois at Urbana-Champaign. The purpose of this research is to identify the characteristics of a standardized training for EPT raters and to develop and evaluate an efficient training model which can improve raters' skills required in oral and essay tests. You have two options; you can participate in only the training workshop or post-rating session.

If you agree to participate in the training workshop the research will be incorporated around your ongoing involvement in a series of activities such as training workshop, and post-rating session. We have tried to design the research project so that it will involve only about 3-4 hours of additional time other than your current EPT commitment. Please see the description below and mark all activities you want to participate;

1. [ ] Survey questionnaire will be provided to evaluate whether the current training program is helpful for current EPT raters. The survey will be provided by the researcher and done anywhere. It is anticipated that it takes approximately 1 hour, and will be anonymous.
2. [ ] Training workshop: you will participate in a standardized training session and complete a survey form after the workshop in order to evaluate a new training program. The workshop will last for one day, approximately 3-4 hours and you are asked to fill out survey questionnaire right after the workshop in order to evaluate a new training program. It is anticipated to take approximately 20 minutes to complete it, which will occur in a classroom in the Foreign Language Building. The workshop will be audio-recorded and all data from the workshop activities will be analyzed.
3. [ ] Post-rating sessions for writing test: improvement in reliability will be estimated by comparing rating results before with after a training session. Researcher will provide the essay samples you will score.
    1) Writing raters will score the existing essay samples and write a reflection log (written report) about the reason why you give a score on the essay in pre- and post training

rating sessions. It is anticipated to take approximately 4-5 hours to complete this task, and this task will be conducted at your convenience.

2) After finishing your rating, researcher will have a set of selective debriefing interviews with raters and interview will be audio-recorded if necessary and it will takes 30 minutes.

Also, if you agree to participate in a rater training program, your participantion during the entire workshop session will be audiotaped. Therefore, I am asking your permission to use your audio-recorded workshop participation or selected debriefing interviews for this research.

All the data collected in this research will be kept confidential, and I will use your information and results for my doctoral dissertation, poster or conference presentation, and journal article. A pseudonym and codes will be used in any analysis of the data in the final research paper and class discussion with my committee members.

The benefits to the participants would be the opportunity to receive an enhanced training program for EPT raters and to improve your professional skills an EPT rater acquire. Participants who completed all aspects of the post-rating session 2 will also receive $50.00. The only possibility of risk involved would be slight emotional discomfort and fatigue. There will be no report to your supervisor, and no effect on your employment status. You may withdraw your participation in the study at any point but incomplete participation will not receive $50.00. This participation is completely voluntary across all activities.

You will be given a copy of this consent form if you want. If you have any questions about the research or the results, please feel free to contact So-young Jang by e-mail at sojang@illinois.edu and Prof. Fred Davidson by e-mail at fgd@illinois.edu.

I allow my research activities to be audio- recorded ____YES____ NO

Print name: _____ Date : ____/____/____/

Signature:

Questions: So-young Jang, Graduate student, Dept. Educational Psychology at
University of Illinois-Urbana Champaign, Phone: 217-332-3426 E-mail: sojang@illinois.edu
****

If you have any questions about your rights as a research participant please contact Anne Robertson, Bureau of Educational Research, 217-333-3023, or arobrtsn@ad.illinois.edu or the Institutional Review Board at 217-333-2670 or irb@illinois.edu

Participant Consent Form 2

Dear EPT trainer:

I would like to invite you to participate in a research project that examines the development and evaluation of a standardized rater training program for the ESL Placement Test (EPT) raters. This research is being carried out by So-young Jang for a doctoral dissertation under supervision of Professor Fred Davidson in the department of Linguistics at University of Illinois at Urbana-Champaign. The purpose of this research is to identify the characteristics of a standardized training for EPT raters and to develop and evaluate an efficient training model which can improve in raters' interview skills and rating reliability. If you agree to participate in this research, you will participate in a series of activities.

1. Survey questionnaire will be provided to evaluate the current training program. The survey will be provided by the researcher and done anywhere. It is anticipated that it takes approximately 1 hour.

2. Rater training session

2.1 Prototype essay sampling and training material development: You will participate prototype sampling procedure. You will decide a final score for speaking and essay samples and also participated in revising training materials, which will be used in the rater training workshop. This process is necessary to estimate not only training effectiveness but rater reliability by comparing raters' rating results before with after a training session.

2.2 Training workshop: you will lead and guide a standardized training session. The workshop will be lasted for one day, and will be audio-recorded and analyzed using codes.

Also, if you agree to participate in a rater training program, your participantion during the entire wrokshop session will be audiorecorded.Therefore, I am asking your permission to use your audio-recorded workshop participation for this research.

All the data collected in this research will be kept confidential, and I will use your information and results for my doctoral dissertation, poster or conference presentation, and journal article. A pseudomym or codes will be used in any analysis of the data in the final research paper and class discussion with my committee members.

The benefits to the participants would be the opportunity to designing an enhanced training program for EPT raters and to improve your professional skills. The only possibility of

risk involved would be slight emotional discomfort and fatigue. There will be no report to your supervisor, and no effect on your employment statue. You may withdraw your paricipation in the study at any point. Your participation is completely voluntary.

You will be given a copy of this consent form if you want. If you have any questions about the research or the results, please feel free to contact So-young Jang by e-mail at sojang@illinois.edu and Prof. Fred Davidson by e-mail at fgd@illinois.edu.

I allow my research activities to be audio- recorded ____YES____ NO

Print name: _____ Date : _____/_____/_____/

Signature:

Questions: So-young Jang, Graduate student, Dept. Educational Psychology at
University of Illinois-Urbana Champaign, Phone: 217-332-3426 E-mail: sojang@illinois.edu
****

If you have any questions about your rights as a research participant please contact Anne Robertson, Bureau of Educational Research, 217-333-3023, or arobrtsn@ad.illinois.edu or the Institutional Review Board at 217-333-2670 or irb@illinois.edu

Appendix F

Pre-workshop Survey (Raters)

A. Personal Profile

This section is given you to investigate your background. Read each item carefully and mark the best choice.

1. Gender Male ( ) Female ( )

2. Are you a native speaker or non-native speaker?
i) ( ) Native speaker of English
ii) ( ) Non-native speaker of English
iii) ( ) Bilingual

3. How many years have you taught for ESL students?
i) ( ) New teacher
ii) ( ) 1 year
iii) ( ) 2 years
iv) ( ) 3 years
v) ( ) more than 3years

4. How long have you severed as an EPT rater?
i) ( ) New rater
ii) ( ) 1 year
iii) ( ) 2 years
iv) ( ) 3 years
v) ( ) more than 3years

5. Have you participated in an EPT rater training program, if yes, how many?
( )
i) ( ) No
ii) ( ) Yes, how many ( )

6. Do you have experience as a rater at another institute, if yes, where (which test)?
i) ( ) No (only EPT)
ii) ( ) Yes, where ; TOEFL TOEIC IELTS OTHERS ( )
How many time did you participated in training program at another institute?
( )

B. Evaluation of current training program

This questionnaire aims to get your opinion about the current EPT test, and EPT rater training program. This questionnaire is given you to identify a need for a standardized training program for the EPT raters. For each of the following statements, please indicate your opinion by marking one of the four numbers.

Part 1. General evaluation

1	2	3	4
Strongly disagree	Disagree	Agree	Strongly agree

1) EPT Test

Items	1	2	3	4
I think the number of the EPT rating scale is valid to measure examinees' proficiency level. (e.g. Too low, ESL 113/500, ESL 114/501, ESL 115/Exempt)				
I think EPT assessment criteria reflect ESL writing ability that is being measured.				
I think Procter training is helpful for understanding the EPT test.				
I think EPT scale descriptors overally accurately describe each proficiency level.				
I think EPT scale descriptors for "too low" accurately describe examinees' proficiency level.				
I think EPT scale descriptors for "ESL 113/500" accurately describe examinees' proficiency level.				
I think EPT scale descriptors for "ESL 114/501" accurately describe examinees' proficiency level.				
I think EPT scale descriptors for ESL 115/ Exempt" accurately describe examinees' proficiency level.				
I think "Organization" reflects ESL writing ability that is being measured.				
I think "Content" reflects ESL writing ability that is being measured.				
I think "Grammar and lexical choice" reflects ESL writing ability that is being measured.				
I think "Use of sources" reflects ESL writing ability that is being measured.				
I think "Plagiarism" reflects ESL writing ability that is being measured.				
I think double rating system is necessary for the EPT essay tests.				
I think holistic rating is necessary for the EPT essay test.				
I think analytic rating is necessary for the EPT essay test.				
I think discussion session with peer raters is helpful to decide a final score.				
After the consensus process, I frequently change my first decision.				
I think the third rater system is required to solve discrepancies between the first two raters' decisions.				
I think "Globalization" is a topic accurately measuring examinees' writing ability.				
I think "Cloning" is a topic accurately measuring examinees' writing ability.				
I think "Animal Testing" is a topic accurately measuring examinees' writing ability.				
I think the difficulty level of three different topics is equivalent.				
I think my final decision is affected by the EPT essay topic given.				

I think examinees' performance is affected by the different topics.				
The training materials provided relevant information about the different three topics of the EPT.				
I think I have a difficulty how to handle three different topics when rating.				

2) EPT Training program

Items	1	2	3	4
The purpose (e.g. high rater reliability) of the EPT rater training has been achieved.				
The purpose (e.g. high rating accuracy) of the EPT rater training has been achieved.				
The purpose (e.g. high agreement between raters) of the EPT rater training has been achieved.				
The sequence of the program was logically organized.				
The workshop program attended was successfully met for my needs and interests.				
I feel I was ready to conduct interviews/rating an examinee after EPT training session.				
The information was effectively presented to deliver the content using visual/audio aids, and handouts.				
Training materials provided a relevant knowledge of understanding the purpose of EPT tests.				
Training materials provided a relevant knowledge of understanding the rating scales.				
Training materials provided a relevant knowledge of understanding assessment criteria.				
Training materials provided a relevant knowledge of understanding the rating procedures (holistic scoring).				
Training materials provided a relevant knowledge of understanding the consensus process when making a split in essay rating.				
Training focus during lecture session is appropriate to improve the rating skills.				
Prototype samples used during workshop are appropriate for the training workshop.				
The number of prototype samples (12 samples) is sufficient.				
Prototype samples are helpful to understand how to rate essays.				
Individual activities during workshop have allowed me to acquire practical rating skills.				
Group rating activities during workshop have allowed me to acquire practical rating skills.				
Peer feedback during group activity is helpful.				
Trainers' feedback during the workshop is helpful.				
The length (2 hours) of the training workshop was appropriate to train raters.				

After the EPT test, I would like to get some regular feedback on my rating performance.				
I agree that additional training is needed for me to conduct essay rating.				

What part(s) of the EPT training program was the most useful to understand EPT oral interview/ essay rating procedures (Be sure to rank all items).

( ) Lecture ( ) Sample practices ( ) Group work
( ) Peers' feedback ( ) Trainer's feedback ( ) Others
( )

Please, rank all of the topics below in terms of the topic difficulty level you perceive (1-most difficult, 3-least difficult).

( ) Globalization
( ) Cloning
( ) Animal Testing

Which one is the most difficult to handle in rating situation? Please, rank all of the topics below(1-most difficult, 3-least difficult).

( ) Globalization
( ) Cloning
( ) Animal Testing

Part 2. Trainer evaluation

1	2	3	4
Poor	Good	Excellent	Not applicable

Items				
1. Accurate knowledge of content				
2. Delivery skills (lecture, effective examples)				
3. Responsiveness to individuals (feedback)				
4. Responsiveness to group (feedback)				
5. Responsiveness to Q & A				
6. Organization skill of training program				
7. Preparation on workshop materials (e.g. sample practice, visual aids and handouts)				

Part 3 Training needs

Read the following statements, and please indicate to what extent you agree you need training for your improvement.

Categories		No training required	Need some training	Need extensive training	Not applicable
Content	Understanding of the EPT				

knowledge	tests procedures				
	EPT rating scale				
	EPT assessment criteria				
	EPT rating procedures				
	EPT essay topics				
Rating skills	Principles of rating				
	Tips on unratable samples				
	Internalization of proto type samples for each level				
Knowledge of policy	Essay rating focus				
	Consensus process				
Others	Different training method				
	Different training materials				
	(Specify if you have any)				

Part 4. Feedback

Look back on the EPT training program the EPT has carried out and please answer each question.

1. Do you think the EPT essay test accurately measures writing ability?

2. Do you think the EPT rating materials and procedures are appropriate for the scoring process? Could you tell which materials and what activities were helpful for you?

3. What challenges might you face when conducting EPT essay rating?

4. How do you get feedback from a trainer or any staff of the EPT when you have some difficulty in essay scoring? If you want, what kinds of feedback do you want to get?

5. What is one aspect of your essay scoring in which you would most like to improve through the workshop (e.g. lecture, individual or group activities, feedback and so on)?

6. What training material is most needed for your professional improvement?

7. What class were/are you teaching? How did your teaching experience help you in rating essay?

8. What were the strengths and weaknesses of the EPT rater training program?

9. Please share any other comments you have for improvement of the EPT training program?

Thank you for your participation!

Appendix G

Pre-workshop Survey (Trainer)

A. Personal Profile : The purpose of this section is to investigate your background and training experience. Please complete the following.

1. Have you ever taught ESL students?
   a) ( ) No
   b) ( ) Yes

2. How many years have you taught for ESL students?
   a) ( ) I am a new teacher
   b) ( ) 1 year
   c) ( ) 2 years
   d) ( ) 3 years
   e) ( ) more than 3 years

3. How long have you served as an English Placement Test (EPT) rater?
   a) ( ) I have never rated the EPT
   b) ( ) 1 year
   c) ( ) 2 years
   d) ( ) 3 years
   e) ( ) more than 3 years

4. How long have you severed as an EPT trainer?
   a) ( ) I am a new EPT trainer
   b) ( ) 1 year
   c) ( ) 2 years
   d) ( ) 3 years
   e) ( ) more than 3 years

5. Do you have experience as a rater/trainer at another institute?
   i) ( ) No (I have only rated the EPT.)
   ii) ( ) Yes: For which test? TOEFL TOEIC IELTS OTHER
( )

How many times did you participate in training programs at other institutes?
( )

B. Evaluation of current training program: The purpose of this questionnaire is to identify any need for a standardized training program for EPT raters. Parts 1 and 2 concern your understanding of the EPT and your previous experience with the EPT rater training. Please briefly explain each of your answers.

Part 1. Understanding the EPT training program

1. What is/are the primary purpose(s) of EPT rater training?

2. What is your role as an EPT trainer during an EPT rater training session?

3. How do you communicate with EPT raters when they ask your help? How do you give feedback to them regarding essay rating?

4. What do you think constitutes an expert rater for the EPT? What are the criteria/standards that would be applied to an expert rater?

5. What are the strengths and weaknesses of the EPT rater training program you have in mind (including rating system, rating materials, e.g. scale descriptors, assessment criteria and etc.)?

6. What instructional resource is most necessary for raters' professional improvement?

7. What would be the most important outcome of this rater training workshop?

8. What kind of concerns/challenges do you have as a trainer when preparing and implementing the workshop?

Part 2. Workshop Preparation

Please briefly explain your answer to each question based on your previous experience in training EPT raters.

1.How would you operationalize (implement) a rater training program in general?

Workshop Procedures (program)

2. What is/are the focus(es) of the EPT raining program (e.g. reliability, accuracy or agreement)?

Main focus

3. What are your main focuses when developing training materials? What kind of materials do you use for training?

Training Materials (for a short lecture and practice)

4. What are your concerns about enhancing rater reliability when developing training materials including prototype essays and workshop activities?

Concerns

5. How would you plan to teach training materials and activities in order for raters to better understand the content during workshop?

Activities

6. Please share your idea or suggestions for improving EPT training program.

Appendix H

Training Evaluation Form

A. Personal Profile: Your ID ( )

B. Evaluation of a New Training Program

This questionnaire aims to get your opinion about the new EPT rater training program in which you have participated. For each of the following statements, please indicate your opinion by marking one of the five numbers.

Part 1. General evaluation

1	2	3	4	5
Strongly disagree	Disagree	Agree	Strongly Agree	Not applicable

Overall Evaluation					
1. The overall training program was well organized.					
2. The overall training program was effective for improving my rating skills.					
3. This training program met my expectations.					
4. I was satisfied with all aspects of the training materials.					
5. The training schedule provided sufficient time to cover all of the proposed activities.					
Workshop Program					
6. The goals of the training were clearly defined.					
7. Each session was clearly organized.					
8. The topics covered were relevant to raters' responsibilities.					
9. The lecture session was helpful for understanding the rating skills.					
10. The individual practice session was helpful.					
11. The group discussion was helpful.					
12. The prototype essay samples used in the workshop were appropriate for enhancing my understanding.					
13. The number of prototype samples used in the workshop was sufficient for enhancing my understanding.					
14. The feedback from the peer raters who participated in this training is helpful for improving the quality of my rating.					
15. The pace of the training was appropriate for the topics covered.					
16. The time for the lecture session was sufficient for understanding the EPT rating system.					
17. The time for the practice session was sufficient to complete all activities.					
18. The times for the feedback and discussion sessions were sufficient for enhancing my rating skills.					
19. Appropriate aids (e.g. audio-visual) for effective delivery were used.					
20. The materials (e.g. handouts for activities) provided were helpful for understanding the content.					

21. Analytic scoring method was helpful for improving the quality of my rating.					
22. Holistic scoring method was helpful for improving the quality of my rating.					
Learning					
23. This workshop helped me increase my professional knowledge (related to essay rating).					
24. The materials were presented at the right level.					
25. I have learned how to internalize the basic concepts of the EPT levels.					
26. I learned problem-solving techniques for essay rating during the discussion session.					
27. Sufficient opportunity for interactive participation was provided in order to share different perspectives and experiences with peer raters.					
Application					
28. The workshop provided balanced integration between content and practice.					
29. Activity I was helpful for enhancing the quality of my essay rating.					
30. Activity II was helpful for enhancing the quality of my essay rating.					
31. Activity III was helpful for enhancing the quality of my essay rating.					
Motivation					
32. Most of my questions were answered during the training.					
33. I will be able to put what I have learned in this workshop into practice.					
34. I would definitely participate in a future rater training program.					
35. I would recommend this workshop to other EPT raters who did not participate in this workshop.					
Instructor Evaluation					
36. The trainer organized the overall procedures well.					
37. The trainer was knowledgeable about the workshop topic.					
38. The trainer was well prepared for the practice session.					
39. The trainer was well prepared for the discussion session.					
40. The trainer encouraged interactive participation.					
41. The trainer gave sufficient feedback during practice.					
42. The trainer clearly answered questions to solve scoring difficulties.					
43. The trainer used effective training methods to deliver the training content and practice sessions.					

Part 2: Degree of Change

Please indicate your level of confidence with the following topics both before and after this training.

1	2	3	4
Very Uncertain	Somewhat Uncertain	Somewhat Confident	Very Confident

Topic					
Overall content knowledge covered in this training program.	Before this training				
	After this training				
Ability to apply content knowledge presented in this training program to actual scoring.	Before this training				
	After this training				
My rating skills related to scoring.	Before this training				
	After this training				
My motivation to be involved with scoring.	Before this training				
	After this training				

Part 3: Feedback

1. What one thing have you liked most about the workshop so far (in comparison with the previous workshop)? Please explain below.

2. What kind of content or activities would you like to see added to this training? Please explain below.

3. Any other suggestions?

Thank you for participating in this raters' workshop!!

Appendix I

Analytic Scoring Guide for EPT/Diagnostic

Descriptors	Absent 1-2	Developing 3-4	Adequate 5-6	Developed 7-8	Advanced 9-10	Exempt 11-12
FOCUS Degree to which main idea/theme and point of view is clear and maintained.	Absent; unclear; insufficient writing to ascertain maintenance	Attempted; main point unclear or shifts; resembles brainstorming; insufficient writing to sustain issue, may lack controlling idea/thesis statement (or it is flawed)	Subject clear/position is not; infers two or more positions without unifying thesis statement (though thesis may be present); abrupt ending	Bare bones; position clear; main point(s) clear and maintained; Thesis present but may be flawed; launch into support without preview	Position announced; points generally previewed; has an introduction, body and conclusion; clear thesis, appropriately placed	All main points are specified and maintained; effective closing; Purpose clear
SUPPORT/ELABORATION Degree to which main point/elements are elaborated and/or explained by specific evidence and detailed reasons	No support; insufficient writing; Content marked by inaccuracies of source information, OR content is completely off-topic, OR majority of essay is copied	Support attempted; ambiguous/confusing; unrelated list; Summarizes/restates sources rather than uses them to support ideas; May indicate misunderstanding of source material; some overt plagiarism	Some points elaborated; most general/some questionable; may be a list of related specifics; Use of oral and written sources demonstrates basic understanding; Attempts at paraphrase are generally unskillful and inaccurate, Covert plagiarism	Some second-order elaboration; sufficiency okay but not much depth, Attempts to use sources to advance the thesis; evidence of some synthesis of ideas; Moderately successful paraphrase; may contain *isolated* instances of direct copying; may not cite sources OR may cite them incorrectly	Most points elaborated by second-order or more, Summary of source content may contain minor inaccuracies, but good understanding is indicated Sources are cited, though possibly inaccurately	All major points elaborated with specific second order support; balanced/evenness
ORGANIZATION Degree to which logical flow of ideas and text plan are clear and connected	No plan; insufficient writing to ascertain maintenance	Attempted; plan can be inferred; no evidence of paragraphing; Confusion prevails; insufficient writing	Plan noticeable; inappropriate paragraphing; major digressions; evidence of logical	Plan is evident; minor digressions; some cohesion and coherence from relating to topic, Paragraph structure *generally* mastered,	Plan is clear; most points logically connected; coherence and cohesion demonstrated; most points appropriately	All points logically connected and signaled with transitions and/or other cohesive devices; all appropriately

			sequencing	*generally* cohesive	paragraphed	paragraphed; no digressions
CONVENTIONS Degree to which student has mastered grammatical and lexical aspects of English	Problems with sentence construction; insufficient writing to ascertain maintenance; Grammatical and lexical errors are severe; no complexity; even simple sentences are flawed	Insufficient writing, Grammatical and lexical errors impede understanding; awkwardness of expression; general inaccuracy of word forms, Little sophistication in vocabulary and linguistic expression; little sentence variety; sentence complexity not mastered	No more than one major error, many minor; sentence construction below mastery but sentence variety is attempted; Inconsistent evidence of some sophistication in sentence variety and complexity	Minimally developed; no major errors, many minor; mastery of sentence construction, Some grammatical/lexical errors; meaning may be occasionally obscured, but essay is still comprehensible	A few minor errors, but no major errors; mastery of sentence construction, Some grammatical/lexical errors, but meaning is not obscured; linguistic expression exhibits some sentence variety and complexity; neither simplistic and awkward nor smooth and sophisticated	No major errors; minimal or no minor errors advanced punctuation with varied sentence structures.
INTEGRATION Evaluation of the paper based on an overall judgment of how effectively the paper as a whole expresses the basic features in order to address the assignment.	Barely deals with topic; does not present most or all features; insufficient writing; length insufficient to evaluate	Attempts to address assignment; some confusion or disjointedness; insufficient writing	Partially developed; some or one feature not developed, but all present; reader inference required, Essay does not flow smoothly; ideas are difficult to follow	Only the essentials present; paper is simple, informative, and clear	Developed paper; each feature evident, but not all equally developed, Some development of thesis; logical sequencing; reasonable use of transitions, Synthesis of ideas	Fully developed paper; all features evident and equally well developed

Although we give essays a single, holistic score, the assessment is based on internalization of the benchmarks, goals and objectives for the ESL Service Courses. Here is a general analytic to holistic conversion table. Remember, though, if a student is borderline, your holistic analysis may trump the analytic cut scores.

Undergraduates
0-15 Too Low
16-30 ESL 113
31-45 ESL 114
46-60 ESL 115

Graduates
0-18 Too Low
19-45 ESL 500
46-56 ESL 501
57-60 Exempt

Appendix J

Raters' Rating Sheet for Workshop Activity

Rater Cover Sheet

Instructions: Please fill out the information. At the end of this training session, gather all of your scoring sheets and your session feedback/evaluation form and staple them together. Thanks!

Name:

Have you rated before? Y / N

If so, how many semesters have you rated? _____

List classes you have taught in the ESL service courses:

Activity #____ Essay # _____

Category	Score	
Focus		Rate the level of difficulty of placing this student (circle one):
Support/Elaboration		Easy Moderate Difficult
Organization		Essay Descriptors (check all that apply):
Conventions		
Integration		___ Borderline (between two levels)
		___ Inconsistent (Very high in areas and low in others)
		___ Below available levels (placement lower than 113/500)
TOTAL		___ Exempt (check if you would exempt 115 or 500)
Analytical Placement		Additional Comments:
Holistic Placement		
GROUP Placement		

Activity #____ Essay # _____

Category	Score	
Focus		Rate the level of difficulty of placing this student (circle one):
Support/Elaboration		Easy Moderate Difficult
Organization		Essay Descriptors (check all that apply):
Conventions		
Integration		___ Borderline (between two levels)
		___ Inconsistent (Very high in areas and low in others)
		___ Below available levels (placement lower than 113/500)
TOTAL		___ Exempt (check if you would exempt 115 or 500)
Analytical Placement		Additional Comments:
Holistic Placement		
GROUP Placement		

Activity #____ Essay # _____

Category	Score	
Focus		Rate the level of difficulty of placing this student (circle one):
Support/Elaboration		Easy Moderate Difficult
Organization		Essay Descriptors (check all that apply):
Conventions		
Integration		___ Borderline (between two levels)
		___ Inconsistent (Very high in areas and low in others)
		___ Below available levels (placement lower than 113/500)
TOTAL		___ Exempt (check if you would exempt 115 or 500)
Analytical Placement		Additional Comments:
Holistic Placement		
GROUP Placement		

Appendix K

Analysis of Open-ended Questions for Pre-workshop Survey

1.Do you think the EPT essay test accurately measures writing ability?

Theme	Responses
	P1: In some ways, the process is long and it is mostly done early morning. Some students might not give their effort; some might not participate during group activities. However these are mostly things us raters can't control.
	P8: Most of the essay tests do. If the question is more difficult to write about for some majors than for others, that might be favor some students even if simply by giving them more time to work on organization- and linguistic-related issues instead of content issues.
Accurate measure as a placement test	P2: I think the EPT essay measures the specific skills that we are looking for to place the students into ESL courses at UIUC.
	P3: This is a single test, so I think the expectation of an accurate measure of writing ability may not be appropriate. The EPT functions well as a placement test. From my teaching experience, there appears to be a good match between students' writing levels and class levels.
Accurate measure	P4: I think it does
	P5: Strong agree
	P6:Yes, I think it's greatly accurate.
	P7: Yes.

2. Do you think the EPT rating materials and procedures are appropriate for the scoring process? Please indicate which materials and what activities were helpful for you.

Theme	Responses
Sample practice	P1: The sample essay helps us raters to gain knowledge on prototype essays however when doing group work, each rater has different perceptions on each essay, which sometimes is confusing.
	P3: Looking at sample essays was definitely helpful. Having a chance to practice rating prior to grading the essays gave me a better understanding of the rating process.
	P4: What we were given especially the samples were very good, However, I think it wasn't enough.
	P6: Samples were helpful.
	P7: Yes. Rating sample essays is most helpful.
Sample practice and group work	P2: The rater training for me was about 2 years ago, I do not accurately remember the specific materials that we used. I do remember reading many sample essays and practicing with other students. The extensive pair work that we did was extremely helpful in learning the EPT process and benchmark.
	P5: They are appropriate. The sample essays to be scored. The peer raters' discussion of the different scores from the same sample essay.
Other materials	P8: The essay is helpful. It would be helpful to also have a video of the lecture that is given to students. Although the lecture is based on the standard material, the individuality of lecturers might affect the level and direction of processing of the essay question.

3. What challenges might you face when conducting EPT essay rating?

Theme	Responses
Different levels of different components	P1: When an essay has several elements of different levels. Sometimes an essay's content might be a certain level while grammatically speaking might be in another level.
Borderline essays and feedback from instructors	P2: Sometimes, there are borderline essays that raters disagree on because different raters focus on different aspects of the essay make their final decisions. However, the discussion that we have to solve the discrepancy is very helpful. In addition, feedback from instructors of the sections helps the rating process because the instructors know what to expect from students in their classes.
Sufficient training	P3: I was not provided sufficient rater training, so I was not sure what I was expected to do when I first started to rate essays.
Tiredness	P4: Getting tired from all the reading, and thus by the end it becomes harder to distinguish between good and bad essays.
Absence of teaching experience and training/peer feedback needs	P8: If the rater has taught one or more writing sequences in the Writing Courses, the benchmarks are made more tangible by the practice of dealing with those students in the day-to-day practice. In the absence of such experience, training must be more elaborate and peer-feedback carries a special role.
Rater's preference	P7: Raters bring in their own preference in rating.
No challenges	P5: No, I don't think there are any challenges up till now.
	P6: None

4. How do you get feedback from the EPT trainer or staff when you have some difficulty in essay scoring? If you want feedback, what kinds of feedback do you want to get?

Theme	Responses
Peer raters	P1: Usually we would discuss the score within the raters, and most of the time comes up with a consensus.
	P2: The raters might ask the ESL coordinator if there is an essay that is extremely difficult to grade because raters cannot come to a consensus. If a third or a fourth rater cannot make the decision, the essay would go to the trainer, but that rarely happens.
	P3: I was not aware that we could get feedback from the trainer or staff. This information was not given at the time of the rating. The only thing I remember in regards to feedback was that, if there was an essay that raters could not come upon an agreement with we should give it to the coordinator. I would like to get feedback on the consensus process. Why there were differences on the ratings.
	P5: Through discussion. Oral discussion
Trainer	P4: Perhaps looking at a specific paragraph with the trainer and saying out loud what are the strengths and weaknesses of that

	paragraph.
	P7: I went to Ann for help. She is a very experienced rater. Feedback that can help me in understanding the relationship between the rating criteria and the curriculum of ESL courses is most helpful.
	P8: I usually take the concrete essay to the EPT trainer and ask for feedback on the issue that I am having difficulty with. For instance, if I am not sure that the person has used the sources sufficiently, I take the article and the essay to the EPT trainer, tell her why I am finding it difficult to evaluate that point, and get feedback. I prefer concrete contextualized feedback.
Personally	P6: Personally.

5. What is one aspect of your essay scoring in which you would most like to improve through the workshop?

Theme	Responses
Essay levels and borderline essay	P1: It's always challenging when rating "tweeners". A detailed instruction on that would be helpful.
	P3: The trainer should be more experienced with rating. He/She should have a clear idea of the levels, the purpose of the test, the meaning of the criteria. Also, raters should be given more time to practice rating before they go through the rating process. Inter-rater reliability needs to be checked even for the practice sessions.
Difficulty in rating the assessment criteria	P2: It is sometimes difficult to place a student because of the criteria "use of sources." A student might have a well-organized essay with good language use, but if the student failed to use sources, it is sometimes confusing where to place the student.
	P4: Looking more on the sentence level. Also, thinking more about the content and what is expected in each level in terms of content.
	P8: Organizational issues for different kinds of essays at different levels. What is a good argument for a 113 student is not good enough, from the point of view of organization, for a 114 student. Often it is linguistic issues that separate those levels. I would like to be able to use organization as a discriminatory criterion to the extent that I use the other criteria.
Hand-writing	P5: The ability to read students' hand-writing.
None	P6: None.
	P7: None.

6. What training material is most necessary for your professional improvement?

Theme	Responses
Samples and feedback	P1: More examples!
	P2: Actual practices using samples was the most useful activity in the training.
	P3: Exposure to more sample essays. Feedback from a rating expert.
	P4: Looking at paragraphs rather than essays, and saying their strengths and weaknesses.

	P8: Essay rating samples.
Scoring criteria	P5: The criteria of the score level.
None	P6: None
	P7: None.

7. What class were/are you teaching? How did your teaching experience help you in rating essays?

Theme	Responses
Rating standards	P1: ESL 114, 115. Actually this is my first semester teaching ESL 115 so my standard would be to find an ESL 114 level essay and rate lower or higher.
	P2: ESL 501. The training took place before the spring semester, so I already had one semester of teaching experience. The teaching experience was very helpful because I had read the student essays in ESL 501 and knew what I should look for in grading.
	P3: ESL 113/500. I had a clear idea of the writing levels for the classes I taught.
	P4: ESL114 and ESL115. Through teaching and giving feedback I am now more aware of what is expected in these two levels.
	P5: ESL 501. Help me understand students' academic writing abilities not only basing on the criteria given but also on the teaching experience.
	P6: I was teacher ESL 500. It helped me a lot on rating. I knew what to look for in EPT teacher's papers.
	P7: ESL 114, 115 and 500. Yes.
	P8: I have taught all the sequences of RSL writing classes. Among them, I have taught ESL 113, 115, and 500 for one semester each – to the best of y memory, and the rest multiple times. As I have mentioned earlier, upon teaching a specific level of students teachers develop feel for what level an ESL students writes about. That experience helps to complement the training for EPT rating.

8. What were the strengths and weaknesses of the EPT rater training program?

Responses	
Strengthens	Weakness
P1: We could spend more time looking at the details of each sample essay.	
P2: Strengths: good sample practices.	Weakness: insufficient information on the hierarchy of criteria—sometimes it is difficult to borderline essays.
P3: Strength: Easy to understand. Accessible for teachers who have no prior experience in rating.	Weakness: Not well organized. Reliability/Validity of the rating is not guaranteed. Unprofessional.
P8: The EPT rater training programs I have received have been of a high level.	

I have mentioned the strength in the respective part of the survey.	
P5: Help raters how to rate EPT score effectively.	No weaknesses
P6: Strengths: feedback from trainer.	Weaknesses: very few samples, no discussion of personal problem when rating.
P4: N/A	
P7: None	

9. Please share any other comments you have for the improvement of the EPT training program.

Theme	Responses
None	P1: N/A
	P2: N/A
	P5: I think it is good.
	P7: None
Taking related course works and extensive training	P3: If possible, raters should at least take the Language Testing class before they participate in the EPT training program. Since the trainer is not a professional in rater training, the quality of the raters could be improved by having background knowledge in language testing. After going through the training process for EPT, I was really concerned of the quality of the raters. The short 30 min training procedure did not really help much in terms of rating. In my case, I relied a lot on my previous rating experience when rating the EPTs. I was grateful I had other experiences in rating essays otherwise; I would have been totally lost.
Extensive training, more information and more feedback + Refresher courses for EPT raters	P4: It is a good start to help raters know what is expected from them, but it does not give us enough hands on experience before starting to rate the papers.
	P6: Better illustration (explanation) of scoring rubric. It mentioned it's interesting to organize a reason with experienced raters who might share their experiences.
	P8: Longer training programs for inexperienced raters/teachers might be helpful. Much as those raters benefit from the feedback of their more experienced peers, the training may be divided into parts for the novice raters alone, in which they proceed at a slower pace through the practice rating samples and reflect on their choices. The following session, they might work together with the experiences raters to gain perspectives that may come from teaching the writing class. Refresher courses for EPT raters are also a good idea for raters who have not taught or rated in some time and come back to do that.

Appendix L

Comments on the New Analytic Descriptors From the Raters

	Rater Comments
G2	1. Support – We suggest it includes a mention of the student using his or her own experiences or ideas as support, and not only the sources provided. We found, in one of the essays, that the student used his own ideas to support his claims, but did not cite any of the provided sources. While this is not exactly what we expect for the ESL students, this kind of "support" is not mentioned in the rubric. 2. Organization – The differences between absent, developing, and adequate aren't very clear, and sometimes we had difficulty judging where to place a student. Even if the grammar was below acceptable, or sources were not cited, at times one student still demonstrated paragraphing (on a low level) – it was just hard to reconcile where to place him. 3. Cut off scores – it also seemed like the score range for 115 was pretty huge; perhaps you could increase the 114 score a little more? Not sure, as the 114 score seemed pretty acceptable.
G3	The categories are well developed and easy to understand. The score range is somewhat confusing. Consider collapsing absent and developing, adequate and developed, advanced and exempt. If you have too many ranges it will confuse the raters. I like the idea of using an analytic rubric, but one practical concern is the cost effectiveness of this method. Using specific evidence to support one's decision makes the rating more valid and reliable. Yet, does it really make a difference compared to the previous holistic rating. Is it really worth the time and effort to do this? If you ask raters to use the analytic rubric I expect they will take more time doing the rating. Is the University willing to pay the extra money if the final judgment is not that different? Compare the time spent using both types of scoring systems and what difference it made on the final judgment.
G 4	1. It would be better if there is only one score in one descriptor to reduce confusions when rating. It is somewhat hard for raters to decide only two numbers. 2. Nagham points out that it would be more flexible if descriptors have wider ranges such as 1-5 not 1-2. 3. As for the integration, it is somewhat vague in terms of analytical evaluation since it is similar to holistic scoring. 4. It would be better if there is an independent descriptor focusing on a thesis statement and topic sentences.

Appendix M

Trainer's Evaluation of 29 Representative Essay Samples (New Selection)

Essay #	Level	Topics	Holistic
1	Undergraduate	Cloning	114
2	Undergraduate	Cloning	115
3	Undergraduate	Cloning	113
4	Undergraduate	Cloning	114
5	Undergraduate	Cloning	115
6	Graduate	Cloning	500
7	Graduate	Cloning	500
8	Graduate	Cloning	exempt
9	Graduate	Cloning	501
10	Graduate	Cloning	501
11	Graduate	Globalization	exempt
12	Undergraduate	Globalization	115
13	Undergraduate	Globalization	115
14	Undergraduate	Globalization	114
15	Undergraduate	Globalization	114
16	Graduate	Globalization	501
17	Graduate	Globalization	501
18	Graduate	Globalization	500
19	Graduate	Globalization	500
20	Undergraduate	Animal Testing	115
21	Undergraduate	Animal Testing	115
22	Undergraduate	Animal Testing	114

23	Undergraduate	Animal Testing	114
24	Graduate	Animal Testing	501
25	Graduate	Animal Testing	500
26	Graduate	Animal Testing	500
27	Graduate	Animal Testing	500
28	Undergraduate	Animal Testing	113
29	Undergraduate	Animal Testing	113

Appendix N

Workshop Observation 1

Date: Jan 12th Time: 10:00-1:30 Place: FLB G23 Participant: 10 raters			
Phase	Session	Time	Agenda and Activity
Familiarization	Lecture	10:00-10:40	1.the purpose of the EPT test (lecture) 2.Reading three test topics and test direction –source use/citation (lecture + reading) 3. Rating process: goals and objectives of ESL service course 4. Raters read three test topics. Raters asked some questions. 5. Holistic and analytic guidelines
Norming	Practice I	10:40-11:30	1.Three essays were provided for raters. 2.Raters did individual works and then two raters did group activity (share idea) G1:R1,R2 G2:R3,R4 G3:R5,R6 G4:R7,R8 G5:R9,R10 3.Trainer provided some feedback around groups. 4. Participants shared comments and questions.
	Break and lecture	11:30-12:00	1. Short Break 2. Rating process and rating tips (do/don't)
	Practice II	12:00-12:25	1.Ranking sheet
		12:25-12:40	1.Individual work and group work (5 people) 2.Share two different sets per group (4-5 essays) G1:R1,R2,R3,R4,R5 G2:R6,R7,R8,R9,R10 3.Trainer's feedback/discussion
		12:40-12:50	1.Differences between Holistic and analytic scoring 2.Placement problems 3. Consensus procedures

			4.Borderline essays 5.Different score on different criteria 6.Scores below 113 or 500
	Practice III	12:50-1:15	1. Simulated the EPT situation, role of third rater. 2.2-3 essays were provided for two groups. G1:R1,R2,R3,R4,R5 G2:R6,R7,R8,R9,R10 3. Trainer provided holistic score with groups.
Closing	Evaluation	1:15-1:30	1.Raters submitted rating sheet and filled out evaluation form.

Appendix O

Workshop Observation 2

Date: Jan 15th Time: 4:00-6:30 Place: FLB 3050 Participant: 5 raters				
Phase	Session	Time	Agenda	Activity
Familiarization	Lecture	4:20-5:00	1.the purpose of the EPT test (lecture) 2.Reading three test topics and test direction – source use/citation (lecture + reading) 3. Rating process: goals and objectives of ESL service course 4. Holistic and analytic guidelines	Raters read three test topics. Raters asked some questions.
Norming	Practice I	5:00-5:15	1. Three essays were provided for raters. 2. Raters did individual works and then two raters did group activity (share idea) G1:A,B,C,D,E 3. Trainer provided some feedback around groups. 4. Rating process and rating tips (do/don't), borderline essays-go down, default, lower score (diagnostic test) 5. Participants shared comments and questions.	Analytic scoring-use of the bullet point, organization
	Practice II	5:15-5:55	1.Individual work(-5:40) and group work (5 people) 2.Share two different sets per group (4 essays, graduate) 3.Trainer's feedback/discussion, original score comparison	
		5:55-6:00	1.Consensus Process about borderline essays	
	Practice III	6:00-6:10	1. Simulated the EPT situation, role of third rater. 2. 2 essays were provided for two groups. 3. Trainer provided holistic score with groups.	
Closing	Evaluation	6:10-6:30	1. Raters submitted rating sheet. 2. Raters filed out evaluation form.	

Appendix P

Post-workshop Survey Evaluation: Open-ended Questions

Rater	Answer
R1	None
R2	1.I've never taken part in this type of workshop before, but actually rating papers +comparing scores is great.
R3	1. What one thing have you liked most about the workshop so far (in comparison with the previous workshop)? Please explain below. What I liked the most about this workshop was the change made in the evaluation rubric itself. The previous version lists organization as the first criterion, which is followed by content, source use, and vocabulary & style. I really like that the current analytical evaluation indicates "focus and support/elaboration" before organization, convention, integration. This is one of the official changes that I really wanted to see as an instructor as well as an EPT rater for a long time. When I grade students' essays—both native speakers and nonnative speakers—I do pay greater attention to the writer's idea (thesis) development more than anything else. I am glad that this change is officially addressed in the revised evaluation criteria. Second, I also like the active integration of technology into the rater training. All the new resources and materials on the ESLTA blogspot space are very helpful for the TAs in many aspects. Sitting in the room both as a rater and trainee, I strongly felt that the use of technology in the EPT recalibration serves multiple purposes. 2. What kind of content or activities would you like to see added to this training? Please explain below. First of all, I really liked that Susan provided the instructors with a bird-eye view on the ESL composition requirement for undergraduate students at UIUC in comparison with the Rhetoric and Speech Comm. courses. But I was thinking that it would be better if the instructors learned the evaluation criteria of Rhetoric courses as well. According to the UIUC composition requirements, Rhetoric 105 and ESL 115 are supposed to cover the same materials, and we do see students change their sections from ESL to Rhetoric and from Rhetoric to ESL each semester. So, I think that it is useful that ESL instructors have a better understanding of the pedagogical practices of the native speaker courses as well (at least their evaluation practices). I think that this is an important issue for the ESL service courses because the English language is not completely foreign to many international students who are admitted to the university any longer (as we all know, they grow up interacting with the language through a variety of media because of the advancement of technology). Since the very boundary between ESL (English as a second language) and EFL (English as a foreign language) is getting less and less clear, I think that the phenomenon should be reflected in the scope of the ESL TA training as well (although I understand that this raises the very question about separating the issue of composition into the first language component and the second language component, which is perhaps beyond the scope of the training). 3. Any other suggestions? I definitely understand that this kind of "numerical judgment" in combination with the rater's "holistic judgment" can enhance the rater's evaluative abilities. Although I understand the rationale and I "really" like the changes made in the new system (e.g.

<table>
<tr><td></td><td>the criteria “focus and elaboration” come before organization and convention), I found it the analytical scoring guide a little difficult to use. One of the main reasons is the way the scores are ranged. I am wondering if we really need to have the following six categories.
<table>
<tr><td>Absent
1-2</td><td>Developing
3-4</td><td>Adequate
5-6</td><td>Developed
7-8</td><td>Advanced
9-10</td><td>Exempt
11-12</td></tr>
</table>

I find it quite difficult to decide especially between “developing” and “adequate”. I understand that student writing evaluation itself can be quite a subjective matter, so this kind of analytical evaluation system can be very valuable. Having said that, however, to me, the difference between the two statements “the thesis of the essay is developing” and “the thesis of the essay is adequate” is not very clear (I think that the essay which has “a developing thesis” can be evaluated “adequate” as well). I am wondering if we really need to have six categories like that. How about if we have four categories such as absent, adequate, advanced, and exempt?

Also, I had difficulty choosing the numerical values themselves. For example, when I judge that the focus of the essay should be placed between 3 - 4 or 5 - 6, I see myself wondering around thinking that “should I give 5 or 6?” The “or” between the numbers 3 or 4 and 7 or 8 in the six categories caused some difficulty when I evaluated the essays. Do we really need to have the “dash” in each category in the six criteria like that? Can we simply assign a single number without the dash? How about just “1 for Absent” and “2 for Adequate” something like that? After the Activity I, II, III, to be honest, I was a little dizzy wondering about the rationale for assigning such numerical values (with the dashes) in the six categories (I am sorry that I was not feeling very well on the day as I stayed up the night before, so I could not articulate on my point right on the spot with the other TAs…).

Another point that I want to make is the current evaluation system between graduate courses and undergraduate courses. As an experienced writing teacher myself, for a long time, I have been wondering about the very equation reflected in the current pedagogical practices of the ESL program that ESL 500=ESL 114 and ESL 501=ESL 115. As we all know, the actual needs of graduate students are quite different from the needs of undergraduates; thus, their needs should be reflected in the pedagogical practices of the program including the EPT evaluation system. I was a little surprised when I saw the similar numerical values were assigned for both graduate and undergraduate courses. I understand that training MATESL students to teach graduate courses in which doctoral students are enrolled is quite a complex phenomenon itself since it involves institutional as well as instructional practices on various levels. I will not elaborate on this as it can be beyond the scope of this workshop, but I want to point out the importance of embracing the actual needs of graduate students more fully in the pedagogical practices of the program as well as in the TA professional development.</td></tr>
<tr><td>R4</td><td>1. Peer and group discussion.
2. Examples for each descriptor. E.g. focus-advanced.
3. Workshop was longer than expected. Consider shortening it.</td></tr>
</table>

R5	None
R6	1. The additional bases for scoring essays, i.e., the analytic scoring rubric gives more reasons for the rating I give. 2. More practice with the analytical scoring. It may be useful to give more to it, or new element introduced in this TA training.
R7	1. I liked the activities about doing actual practices for scoring real EPT papers. 2. Providing TAs with typical papers for each level, just for giving TAs a big picture about general scoring.
R8	1. More organized more visual aids. 2. Maybe more focus on details. That is maybe look at the paragraph and evaluate it. This could be used as an opening activity before looking at a whole essay.
R9	1.The grading rubric. It provides a more a more accurate placing system than the holistic one.
R10	1.PPT and well organized, rubric 2.None 3. More time for the activities.
R11	1.Peer review and collaboration part (activity 2) because it was a good way to check the reliability of my rating skill.
R12	1. Analytic rubric was the helpful.
R13	1. Simple procedure well explained, and good variety of sample writing. 2. Maybe a four more examples and really low and really high writing examples to c ompare.
R14	None
R15	1. Having an analytical scoring rubric. 2. More time for feedback and activities could be given.

Appendix Q

Benchmarks for EPT

Graduate essays

1. **Too Low(place in ESL 500; identify for tutoring)**

- Length insufficient to evaluate
- No organization of ideas; no cohesion; like a freewrite
- Content marked by inaccuracies of source information, or content is completely off-topic, or majority of essay is copied
- Grammatical and lexical errors are severe; no complexity; even simple sentences are flawed

2. **ESL 500**

- Length may be insufficient to evaluate; may be off-topic
- Elements of essay organization(intro, body and conclusion) may be attempted, but are simplistic and ineffective
- Essay may lack a central controlling idea (no thesis statement, or thesis statement is flawed)
- Essay does not flow smoothly; ideas are difficult to follow
- Development of ideas is insufficient; examples may be inappropriate; logical sequencing may be flawed or incomplete
- Paragraph structure not mastered; lack of main idea (topic sentence), focus and cohesion
- Summarizes/restates sources rather than uses them to support ideas
- May lack synthesis of ideas (of the two sources or of sources and student's own ideas)
- May indicate misunderstanding of source material
- Attempts at paraphrase are generally unskillful and inaccurate
- Some overt plagiarism
- Grammatical and lexical errors impede understanding; awkwardness of expression; general inaccuracy of word forms
- Little sophistication in vocabulary and linguistic expression; little sentence variety; sentence complexity not mastered

3. **ESL 501**

- Length is sufficient for full expression of ideas
- Writes on topic
- Elements of essay organization are clearly present, though they may be flawed
- Attempt to advance a main idea; presence of thesis statement
- Flows somewhat smoothly
- Some development and elaboration of ideas; evidence of logical sequencing; transitions may show some inaccuracies
- Paragraph structure *generally* mastered, *generally* cohesive
- Attempts to use sources to advance the thesis; evidence of some synthesis of ideas
- Use of oral and written sources demonstrates basic understanding
- Covert plagiarism; attempted summary and paraphrase; may contain *isolated* instances of direct copying; may not cite sources, or may cite them incorrectly
- Moderately successful paraphrase in terms of smoothness
- Some grammatical/lexical errors; meaning may be occasionally obscured, but essay is stil

l comprehensible
- Inconsistent evidence of some sophistication in sentence variety and complexity

4. **Exempt**
- Contains an intro, body and conclusion
- Clear thesis statement, appropriately placed
- Good development of thesis; logical sequencing; reasonable use of transitions
- Paragraphs are fairly cohesive
- Good synthesis of ideas
- Summary of source content may contain minor inaccuracies, but good understanding is indicated; effective, skillful paraphrase
- Sources are cited, though possible inaccurately
- May contain minor grammatical/lexical errors, but meaning is clear
- Strong linguistic expression exhibiting academic vocabulary, sentence variety and complexity

Benchmarks for EPT composition scoring: Undergraduate essays

1. **Too Low (place in ESL 113; identify for tutoring)**
- Length insufficient to evaluate
- No organization of ideas; no cohesion; like a freewrite
- Content marked by inaccuracies of source information, or content is completely off-topic, or majority of essay is copied
- Grammatical and lexical errors are severe; no complexity; even simple sentences are flawed

2. **ESL 113**
- Length may be insufficient to evaluate; may be off-topic
- Elements of essay organization(intro, body and conclusion) may be attempted, but are simplistic and ineffective
- Essay may lack a central controlling idea (no thesis statement, or thesis statement is flawed)
- Essay does not flow smoothly; ideas are difficult to follow
- Development of ideas is insufficient; examples may be inappropriate; logical sequencing may be flawed or incomplete
- Paragraph structure not mastered; lack of main idea (topic sentence), focus and cohesion
- Summarizes/restates sources rather than uses them to support ideas
- May lack synthesis of ideas (of the two sources or of sources and student's own ideas)
- May indicate misunderstanding of source material
- Attempts at paraphrase are generally unskillful and inaccurate
- Some overt plagiarism
- Grammatical and lexical errors impede understanding; awkwardness of expression; general inaccuracy of word forms
- Little sophistication in vocabulary and linguistic expression; little sentence variety; sentence complexity not mastered

3. ESL 114

- Length is sufficient for full expression of ideas
- Writes on topic
- Elements of essay organization are clearly present, though they may be flawed
- Attempt to advance a main idea; presence of thesis statement
- Flows somewhat smoothly
- Some development and elaboration of ideas; evidence of logical sequencing; transitions may show some inaccuracies
- Paragraph structure *generally* mastered, *generally* cohesive
- Attempts to use sources to advance the thesis; evidence of some synthesis of ideas
- Use of oral and written sources demonstrates basic understanding
- Covert plagiarism; attempted summary and paraphrase; may contain *isolated* instances of direct copying; may not cite sources, or may cite them incorrectly.
- Moderately successful paraphrase in terms of smoothness
- Some grammatical/lexical errors; meaning may be occasionally obscured, but essay is still comprehensible
- Inconsistent evidence of some sophistication in sentence variety and complexity

4. ESL 115

- Contains an intro, body and conclusion (reasonable attempt)
- Clear thesis statement, appropriately placed
- Some development of thesis; logical sequencing; reasonable use of transitions
- Paragraphs are fairly cohesive
- Synthesis of ideas
- Summary of source content may contain minor inaccuracies, but good understanding is indicated
- Sources are cited, though possible inaccurately
- Some grammatical/lexical errors, but meaning is not obscured; linguistic expression exhibits some sentence variety and complexity; neither simplistic and awkward nor smooth and sophisticated (for ESL 115)

Appendix R

Raters' Reflection Log

The purpose of reflection log is to identify your challenges or difficulties in your decision when you using the EPT holistic or analytic rating criteria and descriptors. If you find out some problems with the descriptors, rating criteria or essay, please, write down your concerns on it.

1.Comments on Essays

Please, give the essay number and could you make comments on how did you reach the scores on essay? Any challenges or difficulty?			
Level (under/grad)	**Topic**	**Essay #**	**Your comments or concerns?**

2.Comments on rating materials

	Overall comments on rating materials after finishing your rating	
Analytic Scores	**Focus**	
	Support/Elaboration	
	Organization	
	Conventions	
	Integration	
Holistic Scale	**Under (TOO LOW)**	
	ESL 113	
	ESL 114	
	ESL 115	
	Grad (TOO LOW)	
	ESL 500	
	ESL 501	
	EXEMPT	
Topics	**Animal Testing**	
	Cloning	
	Globalization	
Others	**Any differences between examinees' degree level (Under vs Grad)**	
	Scoring Method used in this study	

Appendix S

Analysis of Rating Split Rate for 2009

Test Date	Topic	Agreement			Disagreement			
		Level	Frequency	Sub-total	Level	Frequency	Sub-total	
1/15	Globalization	114	1	7	501/ex	1	2	9
		115	1		500/501	1		
		500	1					
		501	2					
		ex	2					
		505/507		4				
1/16	Animal Testing	114	2	10	115/114	2	2	12
		115	0					
		500	1					
		501	7					
1/17	Cloning	114	0	8	114/115	1	6	14
		115	4		500/501	2		
		500	1		501/ex	2		
		501	3		Ex/501	1		
1/24	Globalization	114	9	46	114/113	1	6	52
		115	29		114/115	1		
		500	0		115/114	2		
		501	8		501/ex	1		
					501/500	1		
5/27	Globalization	500	11	37	500/501	2	4	41
		501	17		501/500	1		
		ex	9		501/ex	1		
		No marking		11				
		505		4				
6/10	Globalization	115	1	5			0	5
		500	1					
		501	3					

6/11	Cloning	501	1	1			0	
		No marking	1	1				
		505	1	1				
6/20	Animal Testing	500	1	2				2
		501	1					
		No marking	2	2				
		505	4	4				
8/15	Cloning	114	2	41	113/114	1	19	60
		115	1		114/115	1		
					115/114	1		
		500	20		500/501	4		
		501	13		501/500	10		
		Ex	5		501/Ex	1		
					Ex/501	1		
8/17	Animal Testing	114	3	31	114/115	2	12	43
		115	4		500/501	4		
		500	8		501/500	3		
		501	13		Ex/501	3		
		Ex	2					
		507	1	1				
8/18	Globalization	114	4	70	114/115	2	25	95
		115	2		115/114	2		
		500	48		500/501	16		
		501	16		501/500	4		
		505/507	7	7	501:Ex/500	1		
8/19	Cloning	114	12	28	113/114	1	26	54
		115	5		114/115	2		
		500	5		115/114	7		
		501	6		500/501	13		
					501/ex	1		

					Ex/501	2		
		507	2	2				
8/20	Animal Testing	113	1	76	113/114	2	17	93
		114	36		114/115	6		
		115	18		115/114	1		
		500	17		500/501	4		
		501	4		500/ex	1		
					501/500	1		
					501/ex	2		
8/21	Cloning	113	5	66	113/114	3	36	102
		114	29		114/113	4		
		115	15		114/115	16		
		500	9		115/114	7		
		501	8		500/501	3		
					501/500	2		
		507		1	Ex/501	1		
8/22	Animal Testing	113	6	98	114/113	2	15	113
		114	34		114/115	5		
		115	15		115/114	7		
		500	9		501/ex	1		
		501	6					
		505/507	29	29				
8/29	Cloning	113	1	62	113/114	1	19	81
		114	33		114/113	3		
		115	16		114/115	7		
		500	4		115/114	2		
		501	7		500/501	2		
		Ex	1		501/500	2		
					501/Ex	2		
9/05	Animal	114	10	23	114/115	2	5	28

	Testing	115	6		115/114	3		
		500	4					
		501	3					
		505	1	1				

Appendix T

Analysis of Rate of Disagreement Based on Test Topic

Globalization					Animal Testing					Cloning				
Date	Agree	Disagree	%	Total	Date	Agree	Disagree	%	Total	Date	Agree	Disagree	%	Total
1/15	7	2	22.00	9	1/16	10	2	16.67	12	1/17	8	5	38.46	13
1/24	46	5	9.80	51	6/20	2	0	0	2	6/11	1	0	0	1
5/27	37	4	9.76	41	8/17	29	12	29.27	41	8/15	39	19	32.76	58
6/10	5	0	0	5	8/20	76	15	16.48	91	8/19	28	20	41.67	48
8/18	61	15	19.74	76	8/22	98	14	12.5	112	8/21	66	35	34.65	101
					9/05	22	5	18.52	27	8/29	55	12	17.91	67
Total	156	26	44.29	182		237	48	16.84	285		197	91	31.60	288

Appendix U

Comparisons of Means and SD of Rater Groups

	First, experienced			First, new			Second, experienced			Second, New		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
1.	5	4.00	.00	5	3.80	.45	3	3.33	.58	2	3.50	.71
2	5	3.80	.45	5	3.40	.89	3	3.33	.58	2	3.50	.71
3	5	3.60	.55	5	3.80	.45	3	3.33	.58	2	3.00	.00
4	5	3.40	.55	5	3.60	.55	3	3.00	1.00	2	3.00	.00
5	5	2.2	.84	5	3.60	.89	3	3.00	1.00	2	3.50	.71
6	5	3.60	.55	5	3.80	.45	3	3.33	.58	2	3.50	.71
7	4	3.50	.58	5	3.80	.45	3	3.00	1.00	2	3.50	.71
8	5	3.80	.45	5	3.80	.45	3	4.00	.00	2	4.00	.00
9	5	3.80	.45	5	3.80	.45	3	3.67	.58	2	3.50	.71
10	5	3.60	.55	5	3.80	.45	2	4.00	.00	2	4.00	.00
11	5	3.80	.45	5	3.60	.55	2	2.50	.71	2	4.00	.00
12	5	3.60	.55	5	3.40	.55	3	3.67	.58	2	3.50	.71
13	5	3.60	.55	5	3.20	.84	3	3.33	.58	2	3.50	.71
14	5	3.00	.55	5	3.60	.55	2	3.00	.00	2	4.00	.00
15	5	3.25	.71	5	3.40	.89	3	3.00	1.00	2	3.50	.71
16	4	2.60	.50	5	3.40	.55	3	2.67	.58	2	3.50	.71
17	5	3.40	.55	5	3.20	.84	3	3.00	.00	2	3.50	.71
18	5	4.00	.55	5	3.60	.55	3	3.00	.00	2	3.50	.71
19	5	3.60	.00	5	3.60	.55	2	4.00	.00	2	3.00	.00
20	5	3.60	.55	5	3.60	.55	3	3.67	.58	2	3.00	.00
21	5	3.60	.55	5	3.40	.89	3	2.67	.58	2	3.50	.71

22	5	3.80	.45	5	3.60	.55	3	3.67	.58	2	2.50	.71
23	5	3.60	.55	5	3.40	.89	3	3.00	.00	1	3.00	
24	5	3.80	.45	5	3.80	.45	3	3.00	1.00	2	3.50	.71
25	5	3.40	.55	5	3.40	.55	3	3.33	.58	2	3.50	.71
26	5	3.20	.84	5	3.40	.55	3	2.33	.58	2	3.00	.00
27	5	3.40	.55	5	3.20	.45	3	2.67	.58	2	3.00	.00
28	5	3.60	.55	5	3.60	.55	3	3.33	.58	2	2.50	.71
29	5	3.60	.55	5	3.60	.55	3	3.33	.58	2	3.00	.00
30	5	3.60	.55	5	3.60	.55	3	3.33	.58	2	2.50	.71
31	5	3.40	.89	5	3.60	.89	3	3.33	.58	2	3.50	.71
32	5	3.40	.55	5	3.80	.45	3	3.33	.58	2	4.00	.00
33	4	3.25	.96	5	3.60	.55	3	3.33	.58	2	3.50	.71
34	5	3.40	.89	5	3.20	.84	3	3.33	.58	2	3.50	.71
35	5	4.00	.00	5	3.80	.45	3	3.33	.58	2	3.50	.71
36	5	4.00	.00	5	3.60	.55	3	3.67	.58	2	3.50	.71
37	5	4.00	.00	5	3.80	.45	3	3.67	.58	2	3.50	.71
38	5	4.00	.00	5	3.80	.45	3	3.67	.58	2	3.50	.71
39	5	4.00	.00	5	3.80	.45	3	3.33	.58	2	3.50	.71
40	5	4.00	.00	5	3.60	.55	3	3.67	.58	2	4.00	.00
41	5	3.80	.45	4	3.75	.50	3	3.67	.58	2	3.50	.71
42	5	3.80	.45	5	3.60	.55	3	3.00	.00	2	4.00	.00
43	5	4.00	.00	5	3.60	.55	3	3.33	.58	2	3.50	.71

Appendix V

Rating Split Rate for Spring 2010

Test	Topic	Agreement		Disagreement	
		Level	Frequency	Level	Frequency
1/14/10	Cloning	114	3	Ex/501	1
		115	8	500/501	3
		500	5	501/ex	1
		501	1		
1/15	Animal Testing	114	2	501/500	1
		115	1		
		500			
		501	2		
		505	2		
1/16	Globalization	114	9	114/113	1
		115	5	114/115	1
		500	4	115/114	3
		501	2	500/501	1
		Exempt	1	501/500	2
1/23	Globalization	114	13	114/115	7
		115	12	115/114	1
		500	6		
		501	4		
		Exempt	1		
		Total	81		22

Appendix W

Analysis of Comparisons of Raw Scores in Post-rating

#	Level	Topic	Final Score	First Rater	Second Rater	R1	R2	R3	R4	A5	A6
1	Under	Animal Testing	115	114/115	115	115	114	114	115	115	114
2			114	114	114	115	114	114	114	114	115
3			114	114	114	114	113	115	115	114	114
4			113	114	113	114	115	114	114	114	114
5			113	113	114	114	114	115	114	115	113
6			114	113	114	114	115	115	114	115	114
7			113	113	113	114	114	115	114	115	115
8			115	115	115	115	115	114	115	115	115
9			115	114	115	114	114	114	114	115	114
10			115	115	115	115	115	114	115	115	113
11			114	115	114	114	114	114	113	113	114
12			113	113	113	114	114	115	114	114	114
13			114	115	114	114	114	114	115	115	114
14			113	113	113/114	114	113	114	113	114	113
15			115	114	115	115	115	115	114	115	115
16		Cloning	113	114	113	114	114	115	114	115	114
17			113	113/114	113	114	114	114	114	114	114
18			114	114	115	115	114	114	113	114	114
19			114	115	114	114	115	115	113	114	115
20			115	115	115	114	115	115	115	115	115
21			114	113	114	114	113	114	114	114	115
22			114	114	114	114	115	114	114	114	115
23			113	114	113	114	114	114	114	114	114
24			114	114	114	114	114	115	114	113	114
25			115	114	115	115	114	114	115	115	114
26			113	113	113	114	114	114	114	113	115
27			113	113	113	115	114	114	114	115	115
28			115	114/115	115	115	115	115	115	115	115

29			115	115	115	115	115	114	115	115	114
30			115	115	114	114	115	114	114	115	115
31		Globalization	113	113	113	114	113	114	114	114	113
32			115	114/115	115	115	114	115	115	115	115
33			113	113/114	114	114	114	113	114	114	115
34			113	113	113	114	114	115	114	115	114
35			114	114	114/115	114	115	114	114	114	114
36			113	114	113	114	114	114	114	114	114
37			113	113	113	115	114	115	114	114	114
38			114	114	114	114	115	114	114	115	115
39			114	114/115	114	115	114	114	114	115	115
40			115	115	114	115	115	114	115	115	115
41	Under	Globalization	115	115	115	114	114	114	114	114	115
42			115	115	115	115	115	114	115	115	115
43			114	114	114	115	114	115	115	115	114
44			115	114	114/115	115	114	114	115	115	115
45			113	113	113	114	114	114	114	114	113
46	Grad	Animal Testing	500	500	500	500	501	500	501	500	500
47			501	501	exempt	501	501	500	501	500	501
48			exempt	exempt	exempt	exempt	500	501	501	exempt	501
49			501	501	501	501	500	500	501	501	501
50			500	501	500	500	500	500	501	501	501
51			501	501/ex	501	500	501	500	501	500	500
52			500	500	501	501	500	500	501	500	501
53			exempt	exempt	exempt	exempt	501	501	501	exempt	501
54			501	500	501	500	501	501	501	501	500
55			501	501	501	501	501	500	501	500	501
56			exempt	exempt	exempt	501	exempt	501	exempt	exempt	501
57			500	501	500	500	500	500	500	500	500
58			500	500	500	501	500	500	500	500	501

59			exempt	501/ex	exempt	exempt	501	501	501	501	501
60			501	501	501	501	501	500	501	500	500
61		Cloning	exempt	exempt	exempt	exempt	exempt	exempt	exempt	exempt	501
62			500	500	500	500	500	500	500	500	500
63			501	501	501	501	501	500	500	exempt	501
64			500	500	500	500	500	low	500	501	501
65			exempt	501/ex	501	500	500	501	500	501	500
66			501	501	501	501	501	501	501	501	500
67			500	500	501	500	501	500	500	500	501
68			exempt	501	exempt	exempt	exempt	501	exempt	Exempt	501
69			500	500	501	500	500	500	500	500	501
70			501	Ex/501	501	501	500	501	501	501	501
71			500	500	500/501	501	501	501	501	501	501
72			501	501	exempt	501	501	501	501	501	501
73			exempt	exempt	501	501	500	500	501	500	500
74			exempt	exempt	exempt	500	501	501	500	501	500
75			501	501	500/501	501	501	500	501	501	501
76		Globalization	exempt	exempt	501	501	501	501	501	501	500
77			500	501	500	500	500	500	501	501	500
78			500	500/501	500	501	500	500	500	500	501
79			501	501	500	500	501	500	500	501	501
80			501	exempt	500	exempt	501	501	501	501	501
81			500	500	500	500	500	500	501	500	500
82			501	501	501	501	501	501	500	501	500
83			exempt	501	exempt	501	501	501	501	500	500
84			exempt	exempt	exempt	exempt	exempt	exempt	501	501	500
85			501	501	501	500	501	501	500	501	501
86			500	500	500	501	500	500	500	500	501
87			501	500	501	501	500	500	500	500	500
88			exempt	501	exempt	501	exempt	500	501	501	501

89			500	500	501	500	501	500	500	500	500
90			exempt	exempt	exempt	501	exempt	501	501	501	500

Appendix X

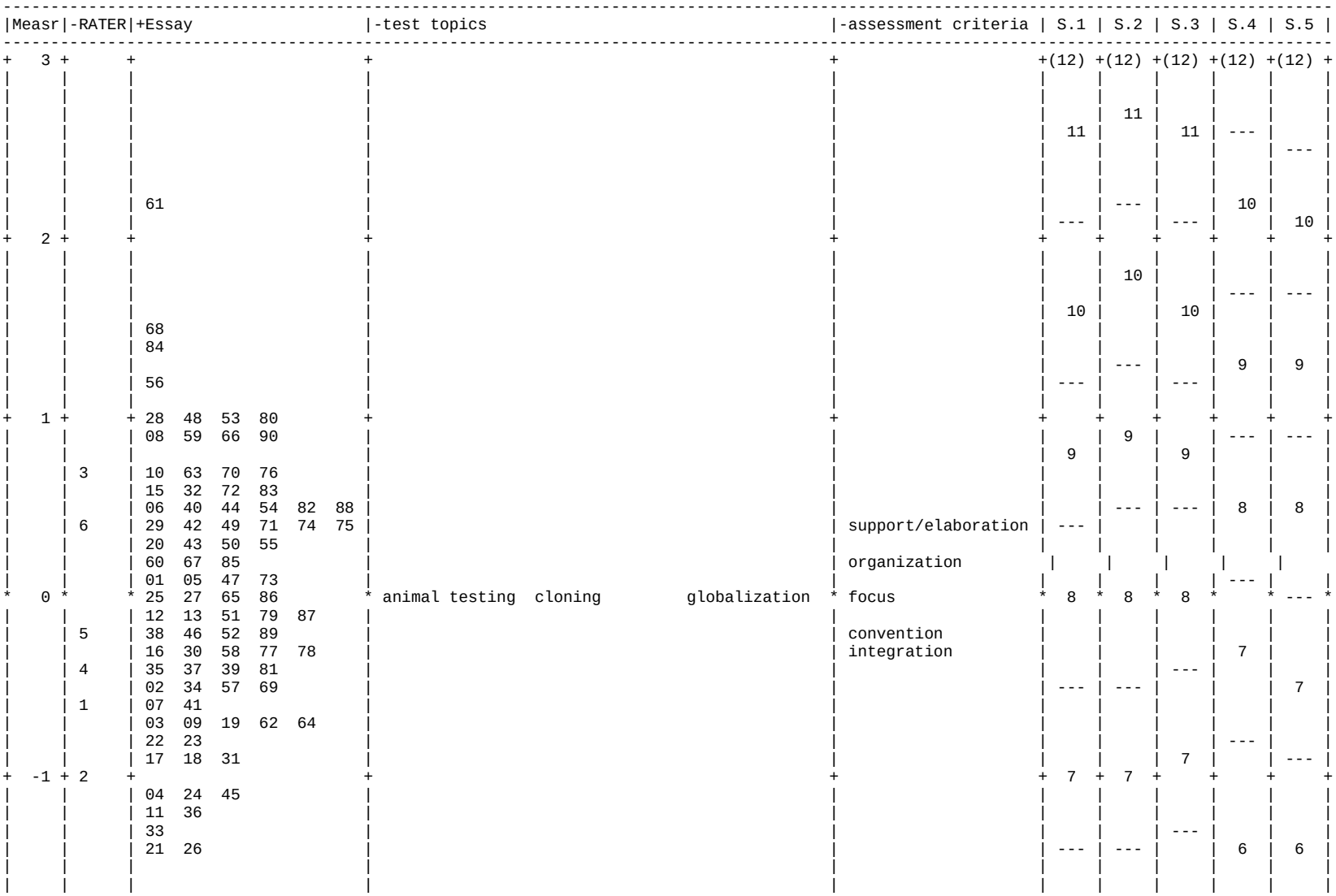
Rating Accuracy Across Language Proficiency Level

Topic	Rater	Method	113	114	115	500	501	Exempt	Total
Animal Testing	R1	Holistic	0	4	4	3	4	3	18
		Analytic	2	5	0	3	4	0	14
	R2	Holistic	1	3	3	4	5	1	17
		Analytic	1	4	3	4	4	0	16
	R3	Holistic	0	3	1	5	1	0	10
		Analytic	1	4	0	5	0	0	10
	R4	Holistic	1	2	3	2	6	1	15
		Analytic	1	4	2	4	4	0	15
	R5	Holistic	0	2	5	4	2	3	16
		Analytic	1	2	3	5	0	0	11
	R6	Holistic	2	4	2	2	3	0	13
		Analytic	0	5	0	5	0	0	10
Cloning	R1	Holistic	0	4	3	4	5	2	18
		Analytic	1	3	0	3	5	0	12
	R2	Holistic	0	2	4	3	4	2	15
		Analytic	0	5	1	4	4	0	14
	R3	Holistic	0	3	2	4	3	1	13
		Analytic	2	4	0	5	0	1	12
	R4	Holistic	0	3	4	4	4	2	17
		Analytic	0	3	0	5	0	0	8
	R5	Holistic	1	4	5	3	4	2	19
		Analytic	1	4	2	5	1	0	13

	R6	Holistic	0	2	3	1	4	0	10
		Analytic	0	5	0	5	0	0	10
Globalization	R1	Holistic	0	2	4	3	2	1	12
		Analytic	1	4	2	3	3	0	13
	R2	Holistic	1	2	2	4	4	3	16
		Analytic	0	2	2	4	3	1	12
	R3	Holistic	2	3	1	5	3	1	15
		Analytic	3	2	0	5	0	0	10
	R4	Holistic	0	3	4	3	1	0	11
		Analytic	0	4	3	5	1	0	13
	R5	Holistic	0	1	4	4	4	0	13
		Analytic	0	4	0	5	0	0	9
	R6	Holistic	2	2	5	3	3	0	15
		Analytic	0	4	0	5	0	0	9

Appendix Y

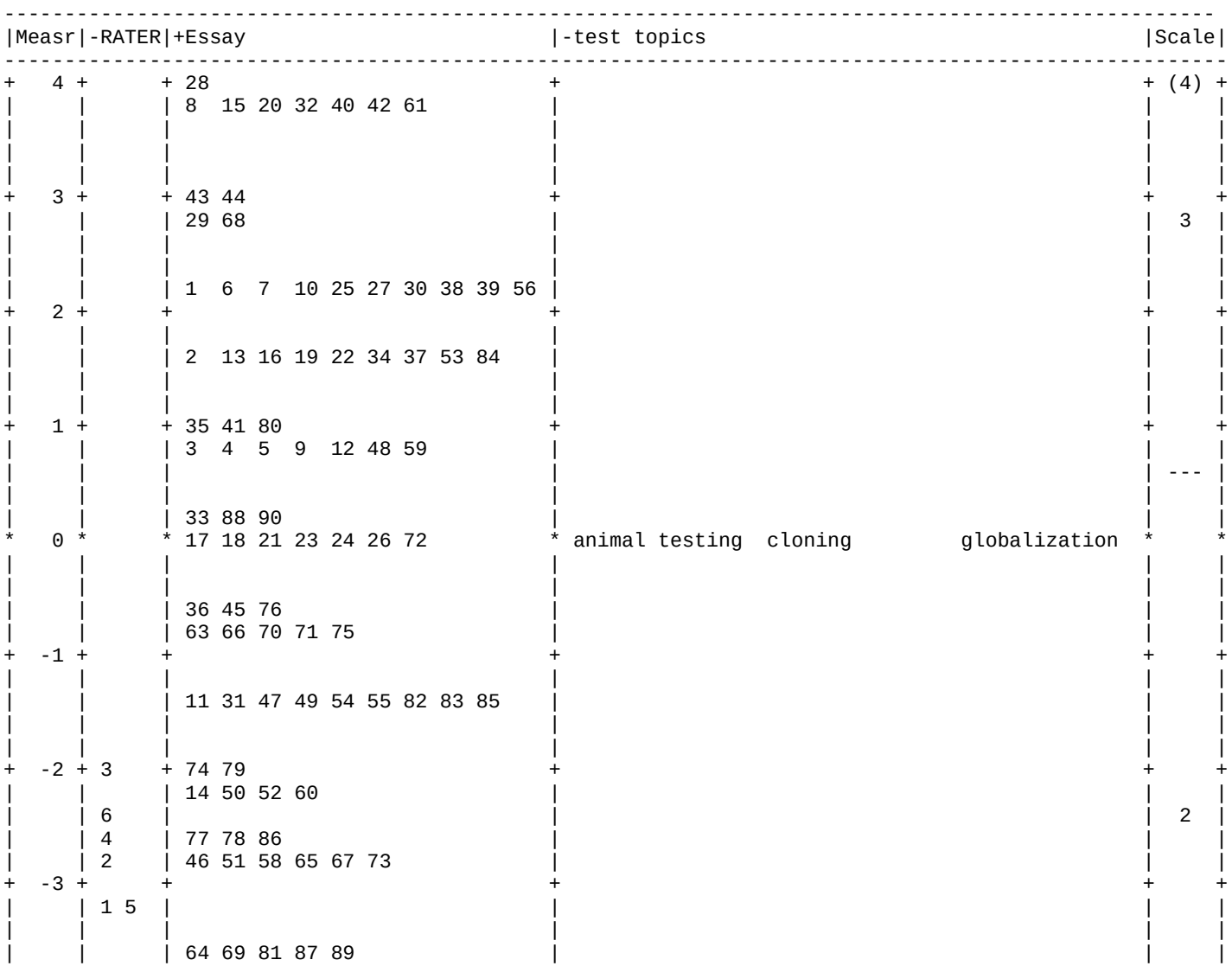
Summary of FACETS Analysis for Analytic Scoring


Measr
-RATER
+Essay
-test topics
-assessment criteria
S.1
S.2
S.3
S.4
S.5
(12)
3
2
1
0
-1
61
68 84
56
28 48 53 80
08 59 66 90
10 63 70 76
15 32 72 83
06 40 44 54 82 88
29 42 49 71 74 75
20 43 50 55
60 67 85
01 05 47 73
25 27 65 86
12 13 51 79 87
38 46 52 89
16 30 58 77 78
35 37 39 81
02 34 57 69
07 41
03 09 19 62 64
22 23
17 18 31
04 24 45
11 36
33
21 26
animal testing
cloning
globalization
support/elaboration
organization
focus
convention
integration

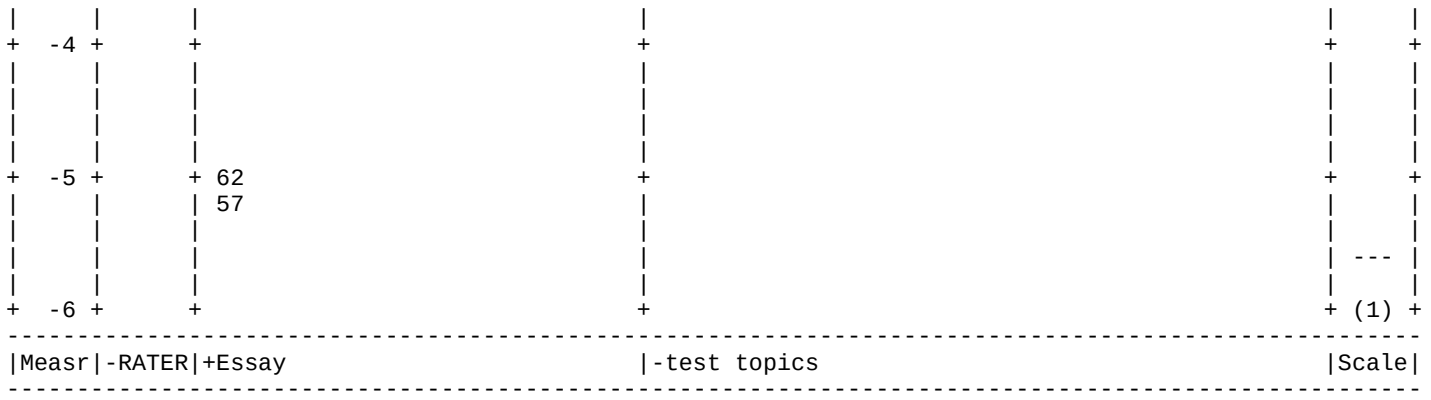

|Measr|-RATER|+Essay |-test topics |-assessment criteria | S.1 | S.2 | S.3 | S.4 | S.5 |
| 14 | (4) | (4) | (4) | (3) | (3) |
-2
6
6
6

Appendix Z

Summary of FACETS Analysis for Holistic Scoring


|Measr|-RATER|+Essay
|-test topics
|Scale|
4
3
2
1
0
-1
-2
-3
28
8 15 20 32 40 42 61
43 44
29 68
1 6 7 10 25 27 30 38 39 56
2 13 16 19 22 34 37 53 84
35 41 80
3 4 5 9 12 48 59
33 88 90
17 18 21 23 24 26 72
36 45 76
63 66 70 71 75
11 31 47 49 54 55 82 83 85
74 79
14 50 52 60
77 78 86
46 51 58 65 67 73
64 69 81 87 89
3
6
4
2
1 5
animal testing
cloning
globalization
(4)
3

2


-4
-5
62
57
-6
---
(1)
Measr
-RATER
+Essay
-test topics
Scale

Appendix AA

Fit Statistics of Essay Used in Post-rating

	Holistic Scoring				Analytic Scoring			
	severity	S.E	Infit	Z	severity	S.E	Infit	Z
1	2.18	.81	.75	-.7	.12	.17	.68	-1.3
2	1.53	.81	.84	-.2	-.46	.17	.95	.0
3	.82	.87	2.40	1.7	-.70	.17	1.92	2.8
4	.82	.87	.66	-.3	-1.12	.17	2.07	3.2
5	.82	.87	2.13	1.5	.09	.17	1.73	2.4
6	2.18	.81	1.04	.2	.55	.17	2.81	4.9
7	2.18	.81	1.11	.4	-.58	.17	1.38	1.3
8	3.80	1.10	.80	.0	.88	.17	1.10	.4
9	.82	.87	.50	-.7	-.67	.17	.80	-.7
10	2.18	.81	1.98	2.6	.69	.17	1.24	.9
11	-1.46	.81	1.03	.2	-1.21	.17	2.51	4.2
12	.82	.87	.85	.0	-.08	.17	1.24	.9
13	1.53	.81	.80	-.3	-.08	.17	.51	-2.2
14	-2.10	.80	.89	-.2	-1.86	.17	1.99	3.1
15	3.80	1.10	.99	.2	.63	.17	.79	-.8
16	1.55	.81	.95	.0	-.34	.17	1.24	.9
17	.05	.91	.03	-2.4	-.94	.17	.80	-.7
18	.05	.91	1.49	.8	-.94	.17	2.08	3.2
19	1.55	.81	2.45	2.5	-.70	.17	2.21	3.4
20	3.82.8	1.10	1.16	.4	.27	.17	.41	-2.9
21	.05	.91	1.74	1.0	-1.42	.17	1.64	2.1

22	1.55	.81	.98	.0	-.76	.17	.96	.0
23	.05	.91	.03	-2.4	-.79	.17	.43	-2.6
24	.05	.91	2.03	1.2	-1.09	.17	1.42	1.5
25	2.20	.81	.75	-.7	.01	.17	.46	-2.5
26	.05	.91	1.89	1.1	-1.42	.17	1.44	1.6
27	2.20	.81	.77	-.7	.01	.17	.81	-.7
28	5.16	1.87			.97	.17	.59	-1.8
29	2.89	.87	.76	-.5	.44	.17	.57	-1.9
30	2.20	.81	.90	-.2	-.31	.17	.32	-3.5
31	-1.37	.81	.86	-.1	-.89	.17	.56	-1.8
32	3.88	1.10	1.05	.3	.60	.17	.64	-1.5
33	.11	.91	1.53	.8	-1.25	.17	1.38	1.3
34	1.62	.81	.95	.0	-.53	.17	.89	-.3
35	.91	.87	.66	-.3	-.41	.17	1.03	.1
36	-.68	.86	.45	-.8	-1.19	.17	1.35	1.3
37	1.62	.81	.97	.0	-.38	.17	1.84	2.6
38	2.27	.81	.90	-.2	-.20	.17	.97	.0
39	2.27	.81	.77	-.7	-.38	.17	.76	-.9
40	3.88	1.10	.80	.0	.46	.17	.97	.0
41	.91	.87	.72	-.2	-.56	.17	.55	-1.9
42	3.88	1.10	.80	.0	.37	.17	.38	-3.1
43	2.96	.87	.99	.1	.32	.17	.49	-2.4
44	2.96	.87	.84	-.3	.54	.17	1.18	.7

45	-.68	.86	.60	-.5	-1.07	.17	.77	-.8
46	-2.78	.86	1.07	.2	-.20	.17	1.63	2.1
47	-1.46	.81	.87	-.1	.15	.17	.95	-.1
48	.82	.87	1.86	1.2	1.05	.17	1.54	1.9
49	-1.46	.81	.72	-.5	.38	.17	.73	-1.0
50	-2.10	.80	.95	.0	.27	.17	.52	-2.2
51	-2.78	.86	1.07	.2	-.14	.17	.66	-1.4
52	-2.10	.80	.97	.0	-.20	.17	.98	.0
53	1.53	.81	.61	-.9	1.02	.17	.67	-1.4
54	-1.46	.81	.99	.1	.52	.17	.63	-1.6
55	-1.46	.81	.87	-.1	.35	.17	.58	-1.8
56	2.18	.81	.89	-.2	1.21	.17	.88	-.4
57	-5.11	1.38	.01	-1.3	-.46	.17	.40	-2.9
58	-2.78	.86	.94	.0	-.26	.17	1.03	.1
59	.82	.87	.51	-.6	.85	.17	.60	-1.7
60	-2.10	.80	.90	-.2	.18	.17	.57	-1.8
61	3.82	1.10	.98	.2	2.22	.18	1.84	2.8
62	-5.09	1.38	.01	-1.3	-.67	.17	.33	-3.3
63	-.74	.86	1.75	1.1	.66	.17	.97	.0
64	-3.63	1.04	2.82	1.9	-.70	.17	1.27	1.0
65	-2.76	.86	1.03	.2	-.02	.17	.73	-1.0
66	-.74	.86	.60	-.5	.86	.17	.66	-1.4
67	-2.76	.86	1.08	.3	.21	.17	.39	-3.0

68	2.89	.87	.76	-.5	1.44	.17	.51	-2.3
69	-3.63	1.04	.99	.2	-.46	.17	.41	-2.7
70	-.74	.86	.66	-.3	.72	.17	.80	-.7
71	-.74	.86	.45	-.8	.41	.17	.69	-1.2
72	.05	.91	.03	-2.4	.61	.17	.36	-3.3
73	-2.76	.86	.92	.0	.10	.17	.87	-.4
74	-2.08	.80	1.03	.2	.41	.17	.66	-1.4
75	-.74	.86	.45	-.8	.38	.17	1.10	.4
76	-.68	.86	.60	-.5	.65	.17	.50	-2.3
77	-2.69	.86	.90	-.1	-.26	.17	.39	-2.9
78	-2.69	.86	.94	.0	-.26	.17	1.07	.3
79	-2.01	.80	.90	-.2	-.08	.17	.68	-1.3
80	.91	.87	.51	-.6	.96	.17	.72	-1.1
81	-3.56	1.04	.98	.2	-.35	.17	.30	-3.6
82	-1.37	.81	.81	-.3	.46	.17	1.19	.7
83	-1.37	.81	1.01	.1	.60	.17	1.13	.5
84	1.62	.81	2.22	2.2	1.43	.17	2.52	4.4
85	-1.37	.81	1.01	.1	.23	.17	1.11	.4
86	-2.69	.86	.94	.0	.03	.17	.86	-.4
87	-3.56	1.04	.78	.0	-.11	.17	1.17	.7
88	.11	.91	1.47	.7	.52	.17	.99	.0
89	-3.56	1.04	.93	.1	-.17	.17	1.72	2.3
90	.11	.91	1.61	.9	.88	.17	.71	-1.2

Appendix BB

Raters' Perception of Rating Difficulty

Essay	Original	R1	R2	R3	R4	R5	R6
1	2	1	1	1	2	2	2
2	1	2	1	2	2	2	2
3	1	2	1	1	3	2	2
4	3	1	1	1	1	2	2
5	3	3	1	1	1	2	2
6	3	3	1	2	1	2	2
7	1	1	1	1	1	2	2
8	1	1	1	2	1	2	2
9	3	2	1	1	2	2	2
10	1	1	1	1	1	2	2
11	3	1	1	1	3	2	2
12	1	2	2	1	2	2	2
13	3	3	1	1	2	2	2
14	2	1	1	1	3	2	2
15	3	1	1	1	2	2	2
16	3	1	1	1	2	2	2
17	2	1	1	1	2	2	2
18	3	2	1	1	2	2	2
19	3	1	2	1	2	2	2
20	1	3	2	1	2	2	2
21	3	1	1	1	2	2	2
22	2	1	1	1	2	2	2

23	3	2	1	1	2	2	2
24	1	1	1	1	2	2	2
25	3	2	1	1	2	2	2
26	1	1	1	1	2	2	2
27	1	1	1	1	2	2	2
28	2	2	1	1	2	2	2
29	1	2	2	2	2	2	2
30	3	3	2	1	2	2	2
31	1	1	1	1	3	2	2
32	2	1	2	1	1	2	2
33	2	1	1	1	2	2	2
34	1	1	1	1	2	2	2
35	2	3	1	1	2	2	2
36	3	1	1	2	2	2	2
37	1	3	1	1	2	2	2
38	1	3	1	1	2	2	2
39	2	2	2	1	2	2	2
40	3	1	1	1	1	2	2
41	1	3	1	1	2	2	2
42	1	2	1	2	2	2	2
43	1	3	1	1	3	2	2
44	2	1	2	1		2	2
45	1	1	1	1	2	2	2

46	1	2	1	1	2	2	2
47	3	3	1	1	2	2	2
48	1	2	1	2	2	2	2
49	1	2	1	1	2	2	2
50	3	3	1	1	2	2	2
51	2	2	1	1	2	2	2
52	3	1	1	1	2	2	2
53	1	2	1	1	2	2	2
54	3	2	1	1	2	2	2
55	1	1	1	1	2	2	2
56	1	1		1	2	2	2
57	3	1	2	1	2	2	2
58	1	2	1	1	2	2	2
59	2	1	1	1	2	2	2
60	1	2	1	1	2	2	2
61	1	3	1	1	1	2	2
62	1	1	1	1	2	2	2
63	1	1	1	1	2	2	2
64	1	1	1	1	3	2	2
65	2	1	1	1	2	2	2
66	1	2	1	1	1	2	2
67	3	3	1	1	2	2	2
68	3	2	1	1	2	2	2

69	3	1	1	1	2	2	2
70	2	2	1	1	2	2	2
71	2	2	1	1	2	2	2
72	3	2	1	1	2	2	2
73	3	2	1	1	1	2	2
74	1	3	2	1	2	2	2
75	2	2	1	1	2	2	2
76	3	2	1	1	2	2	2
77	3	1	1	1	2	2	2
78	2	2	1	1	2	2	2
79	3	1	1	1	2	2	2
80	3	2	1	1	2	2	2
81	1	1	1	1	2	2	2
82	1	2	2	2	2	2	2
83	3	3	2	1	3	2	2
84	1	2	1	1	2	2	2
85	1	1	1	1	2	2	2
86	1	2	1	1	2	2	2
87	3	2	2	1	2	2	2
88	3	2	2	1	2	2	2
89	3	3	1	1	1	2	2
90	1	2	2	1	2	2	2

Appendix CC

Summary of Raters' Reflection Logs for Post- rating

Rater	Comments
R1	• Focus, Organization, Conventions are Required. • Support/Elaboration: I wonder if this descriptor includes the way of doing citations. • Integration: I have found that this descriptor normally corresponds to the holistic score, so it does not play a significant role as an independent descriptor. • Animal Testing: All the topics are controversial, which would be good subjects for argumentative essays. • For undergrads, I have kind of internalized criteria to measure students' writing abilities; however, it was somewhat hard and confusing to evaluate graduates' papers. Also, this directly affects the level of difficulty in rating as well • Scoring method: Raters have not listened to the lecture that students heard, so it would be more helpful to provide raters with the same lecture that students mention in their papers.
R2	• Focus: It would be better if relevant content can be sorted more obviously(one line list one key point)
R3	• ESL 115: The proficiency band for ESL 115 students is wider compared to other levels. Unlike the grads, the undergrads do not have the option of being 'exempt'. Therefore, students who fall in under the exempt category will also be assigned to 115. • Cloning: Sources seem to be outdated. • Examinee: Not sure what this question is about. As a rater, it takes less time to rate undergraduates compared to graduates. Graduate essays are more sophisticated and take more effort from the rater to grade.
R4	• Support/Elaboration: Finding second-order elaboration is one thing, but not all the supporting elements are directly from the articles, but some based on their experiences. Therefore, it was hard to relate it to plagiarism in many cases. In particular, a majority of the "Global Warming" topic essays don't even show an attempt to cite from the article but based on their experiences or from the lecture. • Globalization: As mentioned above, the article of this topic doesn't seem to have enough information compared to the other two topics. I think that's why a lot of the students don't even attempt to cite from the article. The article does not have enough statistics or proof that might support the students' position when writing the essay.
R5	• Any differences between examinees' degree level (Under vs Grad) : Grad's contents are deeper and examples are more persuasive than the undergrads. Grad's grammar, lexicons and their sentence structure are much better than the undergrads • Scoring Method used in this study: Analytic method is helpful for the instructors to know more about their students' writing strengths and weaknesses while holistic method is more efficient to decide at which ESL class level a student should be placed.
R6	• Focus: It is one of the important grading criteria. Students often go off the topic because of their misunderstanding of the topic or over-thinking of the subject. "Focus" can direct raters to sort out essays that went off the topic. • Support: It is certainly necessary to see how well students support their opinions. That makes their essays clear and convincing. • Organization: Also important. The essay has to look well-organized so that students can better show what they think. • Convention: It is hard for students to write grammatically correct essay, but

	convention is important area to assess because raters should assess students' linguistic ability. • Integration: This could be overall impression of the essay, so I think it is good to assess integration. • Cloning: It could be difficult topic. They need enough knowledge to elaborate their opinions. • Examinee level: Grad students seem to know better about writing essays or structures. They can better think in a wider perspective.